Business Case: Netflix - Data Exploration and Visualization

Gyanpriya Misra

DSML-July Beginner Batch

Date: 18/02/2023

Business Case: Netflix - Data Exploration and Visualisation

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

Analyze the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries

---

In [1327]:

```
#importing all libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [1392]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [1328]:

```
#read the Netflix dataset csv file

data = pd.read_csv("netflix.csv")
```

```
In [1329]: #Represent the top 5 rows of dataset

           data.head()
```

Out[1329]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

In [1330]:

```
#basic information about dataset

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

Q1

Defining Problem Statement and Analysing basic metrics:

In 2016, according to “Business Insider” , the major metric for Netflix is economic metric , that is, how much a TV show or movie contributes to Netflix's ability to sign up and retain customers. The main metric is to judge success as a variation on viewing hours called "valued hours." If one particular show is the only thing on Netflix a subset of subscribers watches, but they still pay every month for Netflix, then that show must be really important for them. And Netflix will consider that show valuable to the service.

Netflix majorly focuses on concept of “enjoyment” or how much the show is liked by audience. It segregates the audience in three categories:

1. Starters: “households that watch two minutes of a film or one episode”.
2. Watchers: “households that watch 70% of a film or some episode of a series”.

3. Completers: “households that watch 90% of a film or season of a series”.

```
 Netflix is most popular media and video streaming platforms. But with advancement of time, other
 platforms like Disnep Hotstar,Prime Video and others are also making their place in market. With
 availablility of more options, subscribers base got divided among the platforms. Therefore, it is
 very important to observe other criteria and metrics as well for being in the competition.

 Netflix needs to retain their subscribers and simulatenously attract more customers as well. It is
 the need of the hour to study about the customers, their likes and dislikes, their culture and how
 these factors are affecting the business.

The basic metric which need to be analyzed for growing it's business are:

1. Increase Subscribers base:
   Subscribe base is one of the most important metric. It directly impacts the business revenue and
   the position hold in the market. Netflix needs to attract more customers. As customer base increases,
   data will be more and hence more information leads to better insights and therefore can help in
   taking better decisions.

2. Subscriber retention:
   While attracting new subscribers, it is also important to retain the existing subscribers. The data
   for existing customers leads to some insights which in turn can increase subscribers rate.

3. Popularity of type of contents (world-wide and country-wide):
   It is important to know the popularity of contents for both world-wide and country-wise. Some
   countries prefer Movies rather than TV Shows. However, in some countries, the reverse can be seen.

4. International content growth:
   With availabilty of content in any language, audience are not just stuck with their local
   entertainment, instead, they explore the contents of other countries as well. Some movies
   and TV Shows are so popular that they made their placein international market like "Bahubali"-
   Movie Series of India, "The Stranger Things" - Drama series of America, "Squid Game"-
   Television Series of South Korea and others.
   It is important to consider the type of contents to be release in international market so
   that it can increase the revenue.

5. Know Your Customer:
   Customers or Subscribers are the pillars of any business. It is very important to know about
   your customer, their likes and dislikes. Some like TV Shows while others prefer movies. Some
   prefer action genre while some opt for romantics. On each criteria, it is important to conclude
   that on which category,subscriber base is inclined.

6. Subscriber Enagagement:
   Closely observation on number of hours spending by viewer on particular movie/TV Show. That is,
   how many minutes of movies or how many episodes of first season of any show have been watched.
   This information will convey about the quality of content.

7. Game-changing content:
   Sometimes, subscribers subscribe to platform for watching specific movie/TV Show. This means,
   that movie/TV Show is too important and can brought maximum revenue to the business. There is
   need to closely observe those contents and promote them on international market.

8. Revenue and Marketing:
   It is important to understand how much revenue has been generated from various countries, from
   the types of content added on platform. Metrics such as average revenue per user and lifetime
   customer value are too important for any business. They can drive the business into new directions
   and can help in taking better decisions for pricing and marketing strategy to get maximum revenue.

9. Country-wise analysis:
   Every country and it's people are different. People are closely attached to their culture also.
   Movie which is very popular in one country can be a disaster in other country. Hence, it can
   also be the metric for providing better insights for business growth.
```

Q2

Observations on the shape of data,

In [1331]:

```
data.shape

# The dataset contains 8807 entries of movies/TV shows along with their 12 attributes.
```

Out[1331]:

```
(8807, 12)
```

data types of all the attributes,

In [1332]:

```
data.dtypes

# There are total 12 attributes in dataset.
# The attribute "release_year" is an integer type attribute.
# Remaining attributes are object type.
```

Out[1332]:

```
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object
```

missing value detection,

```
In [1333]: data_missing_value = pd.DataFrame(data.isnull().sum()).reset_index()
           data_missing_value.rename({"index":"Attributes",0:"Count_of_missing_values"},axis = 1,inplace = True)
           data_missing_value

           # There are multiple attributes which have missing values like director,cast, date and
           # release year of movies/shows, rating and duration.
           # The maximum values are missing in director column.
```

```
Out[1333]:
                Attributes  Count_of_missing_values
            0      show_id                        0
            1         type                        0
            2        title                        0
            3     director                     2634
            4         cast                      825
            5      country                      831
            6   date_added                       10
            7 release_year                        0
            8       rating                        4
            9     duration                        3
           10    listed_in                        0
           11  description                        0
```

In [1334]:

```
#Treating date-time columns:

#Filling missing values of date_added column

data["date_added"] = data["date_added"].fillna(data["date_added"].mode()[0])

#Conversion of DataType of column "date_added"

data["date_added"] = pd.to_datetime(data["date_added"])

#As the "date_added" column showing dates but the datatype of this column in dataset is object. Hence, there is
#need of change of datatype of this column.
```

In [1335]:

```
#Creating Year, Month and Day columns

#creating separate columns for year,month and day
data["year"] = data["date_added"].dt.year
data["month"] = data["date_added"].dt.month
data["day"] = data["date_added"].dt.day
data["month_name"] = data["date_added"].dt.month_name()

#dealing with missing values
data["year"] = data["year"].fillna(round(data["year"].mean(),0))
data["month"] = data["month"].fillna(round(data["month"].mean(),0))
data["day"] = data["day"].fillna(round(data["day"].mean(),0))
data["month_name"] = data["month_name"].fillna(data["month_name"].mode()[0])

#changing datatypes of year, month and day
data["year"]=data['year'].astype(int)
data["month"]=data['month'].astype(int)
data["day"]=data['day'].astype(int)
```

In [1336]:

```
#conversion of categorical attributes to 'category' (If required),
#not required
```

statistical summary

In [1337]:

```
data.describe(include = object)

# The count and unique counts of all columns of dataset have been shown below.
```

Out[1337]:

	show_id	type	title	director	cast	country	rating	duration	listed_in	description	month_name
count	8807	8807	8807	6173	7982	7976	8803	8804	8807	8807	8807
unique	8807	2	8807	4528	7692	748	17	220	514	8775	12
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...	July
freq	1	6131	1	19	19	2818	3207	1793	362	4	827

In [1118]:

```
#Observation on wrong entries in dataset.

print("1. Rows having incorrect entries for column 'rating': ")
print()
print(data[data["duration"].isna()][["title","rating","duration"]])
print("___________________________________________________________")
print()
print("2. The unique values of 'rating' columns are: ")
print()
print(data["rating"].unique())

# It can be observed that the rating has been categorized as ['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', '
# TV-Y','TV-Y7','R','TV-G', 'G', 'NC-17', '74 min', '84 min', '66 min', 'NR', nan,'TV-Y7-FV', 'UR'].
# Notice that three values ('74 min', '84 min', '66 min') are not actually a category of ratings.
# Note:There are null values for duration of three movies named "Louis C.K. 2017","Louis C.K.: Hilarious",
# and "Louis C.K.: Live at the Comedy Store". However,the data of duration has been listed under column
# rating for these three movies. There is need of cleaning the dataset.
```

```
1. Rows having incorrect entries for column 'rating':

                                     title  rating duration
5541                       Louis C.K. 2017  74 min      NaN
5794                 Louis C.K.: Hilarious  84 min      NaN
5813  Louis C.K.: Live at the Comedy Store  66 min      NaN
___________________________________________________________

2. The unique values of 'rating' columns are:

['PG-13' 'TV-MA' 'PG' 'TV-14' 'TV-PG' 'TV-Y' 'TV-Y7' 'R' 'TV-G' 'G'
 'NC-17' '74 min' '84 min' '66 min' 'NR' nan 'TV-Y7-FV' 'UR']
```

In [1393]:

```
#Data cleaning based on the above wrong entries.

index_var = data[data["duration"].isna()].index
for i in index_var:
    data["duration"].iloc[index_var] = data["rating"].iloc[index_var]
for i in index_var:
    data["rating"].iloc[index_var] = None
```

In [1339]:

```
#Data has been cleaned

data[data["title"].isin(["Louis C.K. 2017", "Louis C.K.: Hilarious", "Louis C.K.: Live at the Comedy Store"])]
```

Out[1339]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	year	month	da
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	None	74 min	Movies	Louis C.K. muses on religion, eternal love, gi...	2017	4	
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	None	84 min	Movies	Emmy-winning comedy writer Louis C.K. brings h...	2016	9	1
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	None	66 min	Movies	The comic puts his trademark hilarious/thought...	2016	8	1

In [1340]:

```
#Analysis after cleaning of data
data.isna().sum()

#Now, there are no null entries in column "duration" and 7 null entries in column "rating".
```

Out[1340]:

```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added         0
release_year       0
rating             7
duration           0
listed_in          0
description        0
year               0
month              0
day                0
month_name         0
dtype: int64
```

```
In [1341]: #Filling missing values of rating column

            data["rating"] = data["rating"].fillna(data["rating"].mode()[0])

            #Analysis after cleaning of data
            data.isna().sum()
```

```
Out[1341]: show_id            0
           type               0
           title              0
           director        2634
           cast             825
           country          831
           date_added         0
           release_year       0
           rating             0
           duration           0
           listed_in          0
           description        0
           year               0
           month              0
           day                0
           month_name         0
           dtype: int64
```

Q3

Non-Graphical Analysis: Value counts and unique attributes

In [1342]:
```
print("Unique attributes for type of content and their count are: ")
print()
print(data["type"].value_counts())
print("____________________________________________________________")
print()
print("Unique attributes of rating column and their count are: ")
print()
print(data["rating"].value_counts())

#There are total 6131 movies and 2676 TV shows in the dataset.
#There are 14 types of unique ratings in the dataset.
```

```
Unique attributes for type of content and their count are:

Movie      6131
TV Show    2676
Name: type, dtype: int64
```

```
Unique attributes of rating column and their count are:

TV-MA       3214
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
NC-17          3
UR             3
Name: rating, dtype: int64
```

Q4

VISUAL ANALYSIS - Univariate, Bivariate after pre-processing of the data

In [1343]:

```
#Note: Pre-processing involves unnesting of the data in columns like Actor, Director, Country

#Unnesting of Country column
country_unnest=data['country'].apply(lambda x: str(x).split(', ')).tolist()
data_country=pd.DataFrame(country_unnest,index=data['title'])
data_country=data_country.stack()
data_country=pd.DataFrame(data_country)
data_country.reset_index(inplace=True)
data_country=data_country[['title',0]]
data_country.columns=['title','country']
data_country.head()

# data_country DataFrame has been created with countries having unique titles.
```

Out[1343]:

	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan
3	Jailbirds New Orleans	nan
4	Kota Factory	India

In [1344]:

```python
#Unnesting of director column
director_unnest=data['director'].apply(lambda x: str(x).split(', ')).tolist()
data_director=pd.DataFrame(director_unnest,index=data['title'])
data_director=data_director.stack()
data_director=pd.DataFrame(data_director)
data_director.reset_index(inplace=True)
data_director=data_director[['title',0]]
data_director.columns=['title','director']
data_director.head()

# data_director DataFrame has been created with directors having unique titles.
```

Out[1344]:

	title	director
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	nan
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	nan
4	Kota Factory	nan

In [1345]:

```
#Unnesting of cast column
cast_unnest=data['cast'].apply(lambda x: str(x).split(', ')).tolist()
data_cast=pd.DataFrame(cast_unnest,index=data['title'])
data_cast=data_cast.stack()
data_cast=pd.DataFrame(data_cast)
data_cast.reset_index(inplace=True)
data_cast=data_cast[['title',0]]
data_cast.columns=['title','cast']
data_cast.head()

# data_cast DataFrame has been created with cast having unique titles.
```

Out[1345]:

	title	cast
0	Dick Johnson Is Dead	nan
1	Blood & Water	Ama Qamata
2	Blood & Water	Khosi Ngema
3	Blood & Water	Gail Mabalane
4	Blood & Water	Thabang Molaba

In [1346]:

```
#Unnesting of listed_in column
genre_unnest=data['listed_in'].apply(lambda x: str(x).split(', ')).tolist()
data_genre=pd.DataFrame(genre_unnest,index=data['title'])
data_genre=data_genre.stack()
data_genre=pd.DataFrame(data_genre)
data_genre.reset_index(inplace=True)
data_genre=data_genre[['title',0]]
data_genre.columns=['title','genre']
data_genre.head()

# data_cast DataFrame has been created with cast having unique titles.
```

Out[1346]:

	title	genre
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows

In [1347]:

```
#Dealing with missing and NaN/None values in the above four unnested dataset.

#deleting NaN values from director column
index_country = data_country[data_country["country"] == "nan"].index
index_country
data_country.drop(index_country, inplace = True)

#deleting NaN values from director column
index_director = data_director[data_director["director"] == "nan"].index
index_director
data_director.drop(index_director, inplace = True)

#deleting NaN values from cast column
index_cast = data_cast[data_cast["cast"] == "nan"].index
index_cast
data_cast.drop(index_cast, inplace = True)

#filling mode in missing values of genre column
data_genre["genre"] = data_genre["genre"].fillna(data_genre["genre"].mode()[0])
```

BASIC ANALYSIS

In [1348]:

```
#Now we have four more datasets along with the original dataframe.
#They are 'data'(original dataset), 'data_country', 'data_director', 'data_cast', and 'data_genre'.

#1 The first release and latest release in dataset

first_release = data["release_year"].min()
latest_release = data["release_year"].max()
print(first_release ,":", latest_release)

#The first release happened in 1925 and latest release happened in 2021 according to given dataset.
#Alternatively, we can say that this is the time span for which dataset has been provided.
```

```
1925 : 2021
```

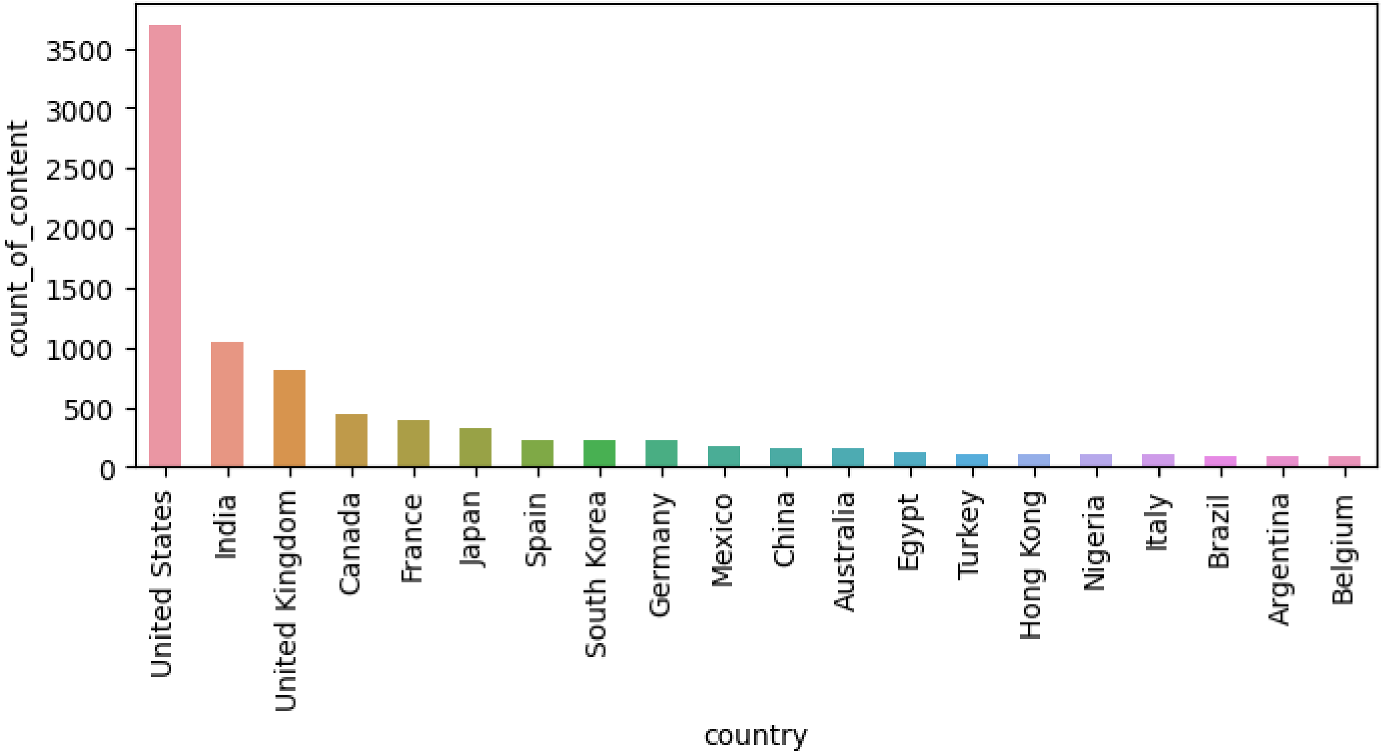
In [1349]:

```
#2 First added date and latest added date of Video content on Netflix Platform.

first_added = data["date_added"].min()
print(first_added)
latest_added = data["date_added"].max()
print(latest_added)

#Netflix has first added their content on Janaury 1st, 2008 and the latest content is added on
#September 25th, 2021 according to dataset
```

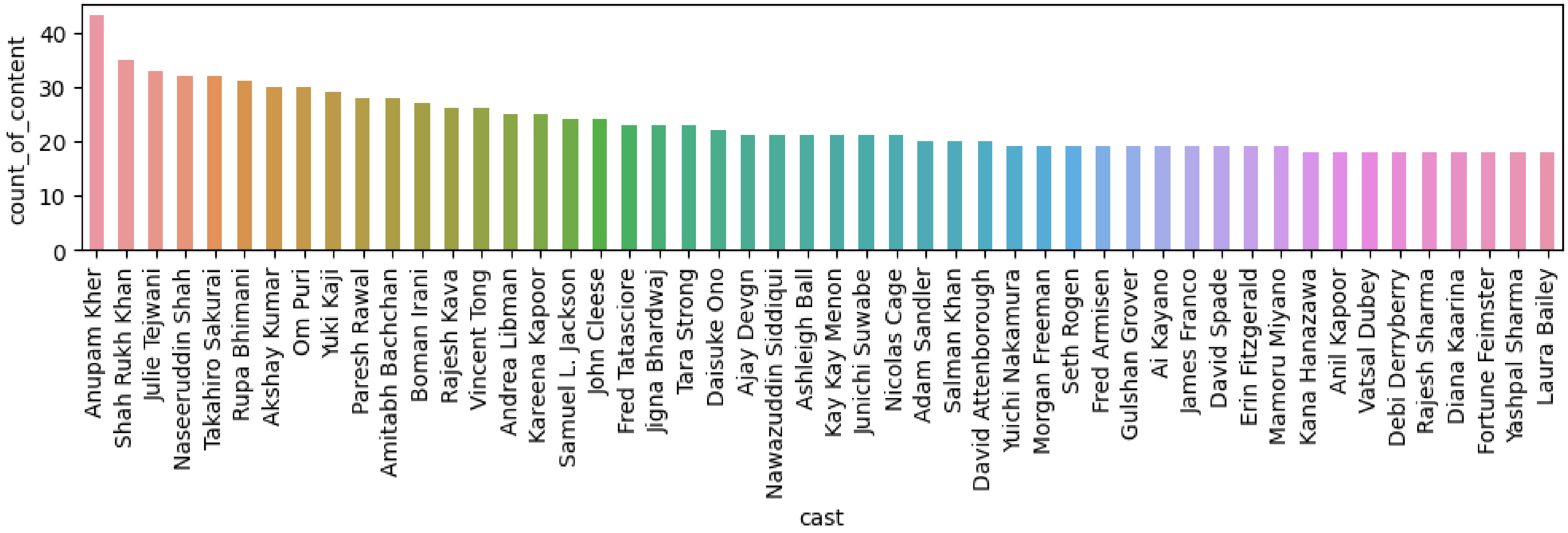
```
2008-01-01 00:00:00
2021-09-25 00:00:00
```

In [1350]:

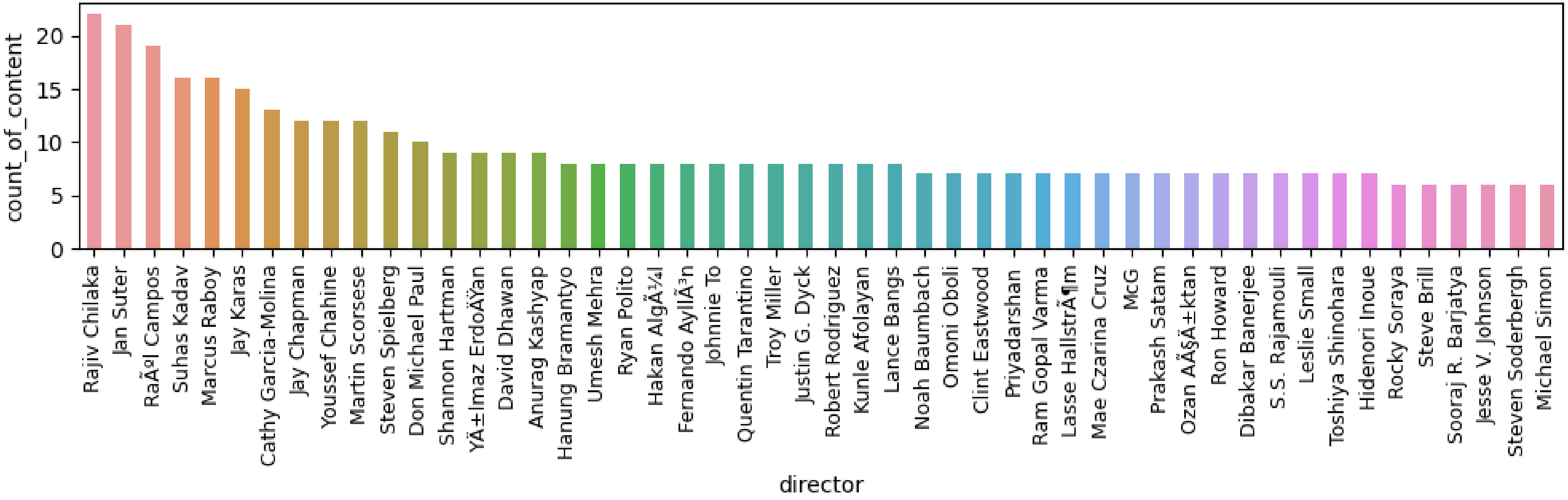
```
#3 Country having maximum releases of content

count_4 = data_country["country"].value_counts()
top_countries = pd.DataFrame(count_4.head(20)).reset_index()
top_countries.rename({"index":"country","country":"count_of_content"},axis = 1,inplace = True)

plt.figure(figsize = (8,3))
sns.barplot(data = top_countries, x = "country", y = "count_of_content", width = 0.5)
plt.xticks(rotation = 90)
plt.show()

#The most releases of content have been happened in United States, followed by India and United Kingdom.
```



```
In [1351]: #4 Most popular cast

count_3 = data_cast["cast"].value_counts()
top_cast = pd.DataFrame(count_3.head(50)).reset_index()
top_cast.rename({"index":"cast","cast":"count_of_content"},axis = 1,inplace = True)

plt.figure(figsize = (12,2))
sns.barplot(data = top_cast, x = "cast", y = "count_of_content", width = 0.5)
plt.xticks(rotation = 90)
plt.show()

#The most popular cast is Anupam Kher with 43 movies/TV shows enacted.
```



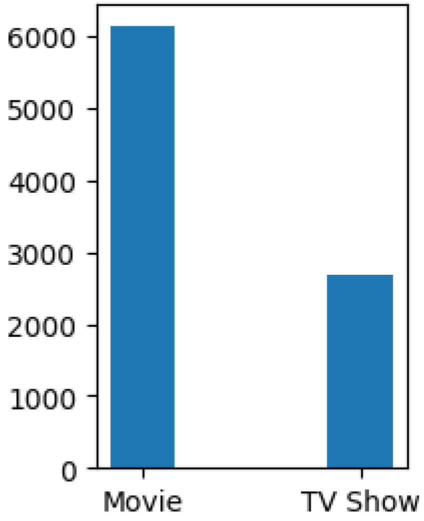
In [1352]:
```
#5 Most popular Director

count_2 = data_director["director"].value_counts()
top_directors = pd.DataFrame(count_2.head(50)).reset_index()
top_directors.rename({"index":"director","director":"count_of_content"},axis = 1,inplace = True)

plt.figure(figsize = (12,2))
sns.barplot(data = top_directors, x = "director", y = "count_of_content", width = 0.5)
plt.xticks(rotation = 90, fontsize = 9)
plt.show()

#The most popular director is Rajiv Chilaka with 22 movies directed.
#Rajiv Chilaka has directed movies of series of "Chota-Bheem", which is very popular among kids.
```



GRAPHICAL : UNIVARIATE ANALYSIS

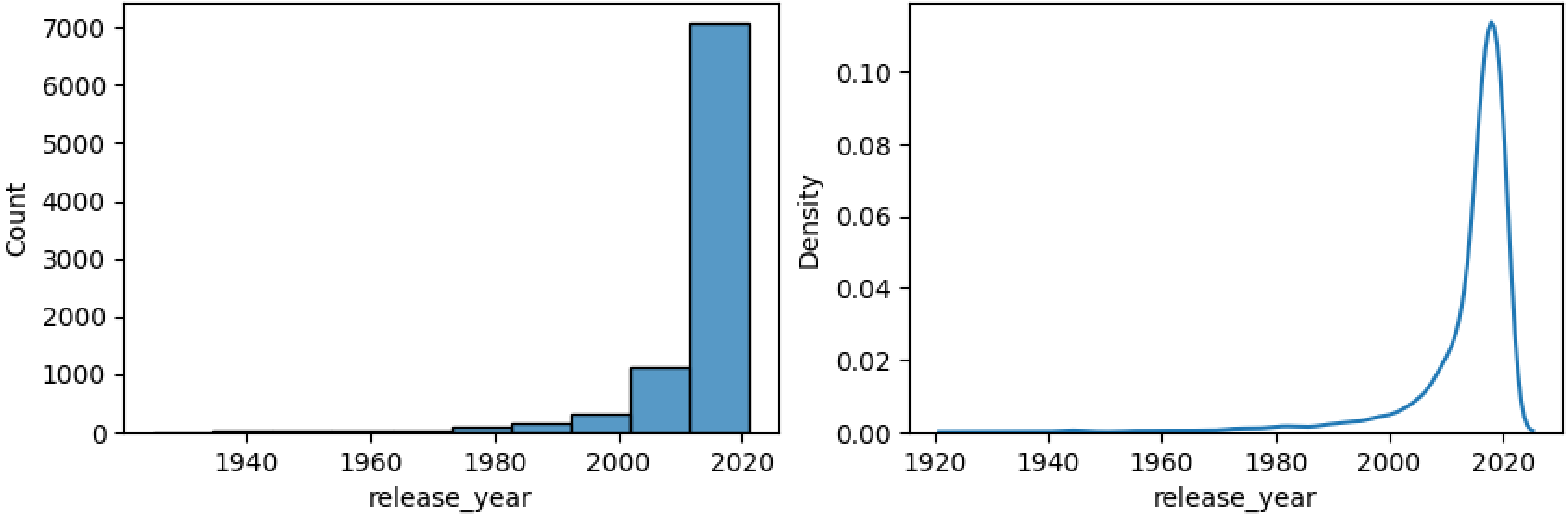
In [1353]:

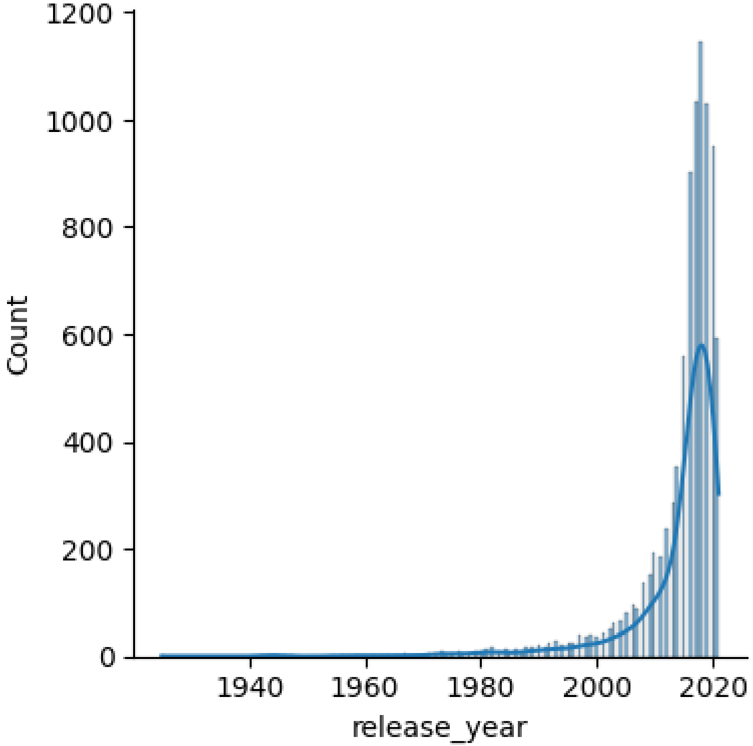
```
# 1. Count of movies and TV shows

plt.figure(figsize = (2,3))
count_1 = data["type"].value_counts()
plt.bar(count_1.index,count_1,width = 0.3)
plt.show()

#The count of movies is far greater than count of TV Shows. People are fond of movies rather than TV shows.
```



In [1354]:

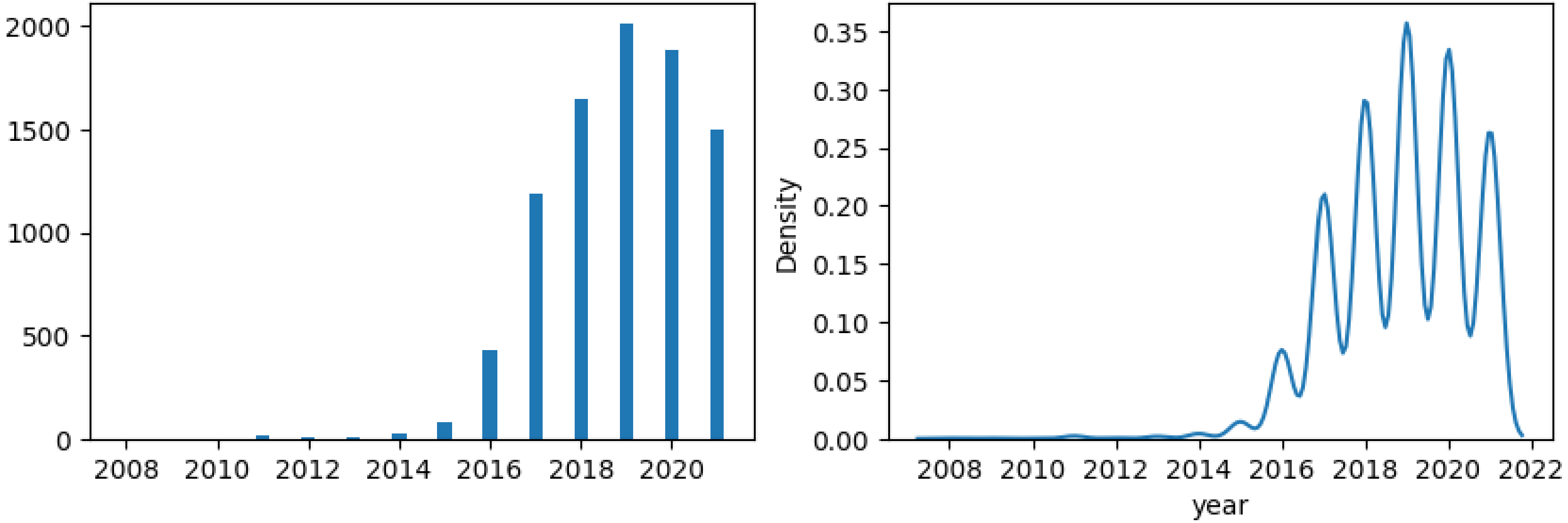
```
# 2. Popularity of Video contents over years

plt.figure(figsize = (10,3))
plt.subplot(1,2,1)
sns.histplot(data["release_year"],bins = 10)
plt.subplot(1,2,2)
sns.kdeplot(data["release_year"])

sns.displot(data["release_year"], kde=True, height=4, aspect=1)
plt.show()

# The cinema has evolved a lot. Earlier, movie making is such a huge task. With advancements of
# technology and enhanced talent,most of people try their hands on production now. It can be
# observe from the plots that after 2000, the production of movies and TV shows have been
# increased drastically.
```




1200
1000
800
600
400
200
0
Count
1940
1960
1980
2000
2020
release_year

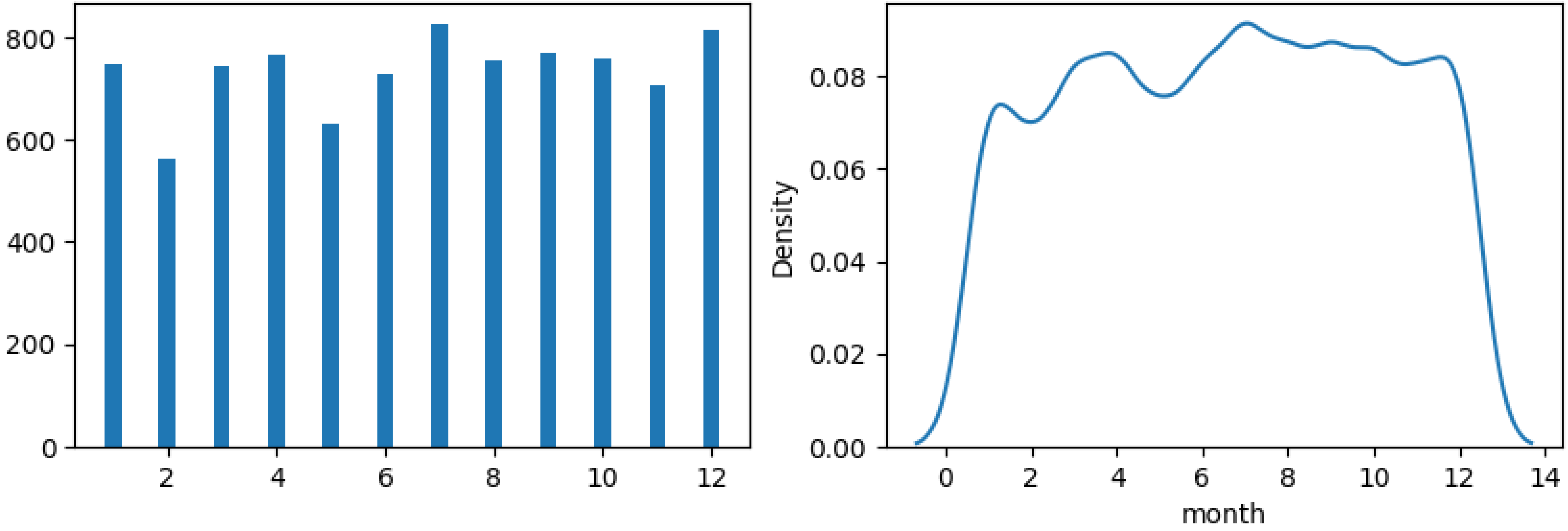
In [1355]:

```
# 3. Video contents over years added On Netflix Platform

plt.figure(figsize = (10,3))
plt.subplot(1,2,1)
count_year = data["year"].value_counts()
plt.bar(count_year.index,count_year,width = 0.3)
plt.subplot(1,2,2)
sns.kdeplot(data["year"])
plt.show()

# Netflix has started adding enormous amount of data from the last five years. From 2016, they have
# raised the quantity in their platform. The maximum movies/TV Shows has been added in year 2018.
```



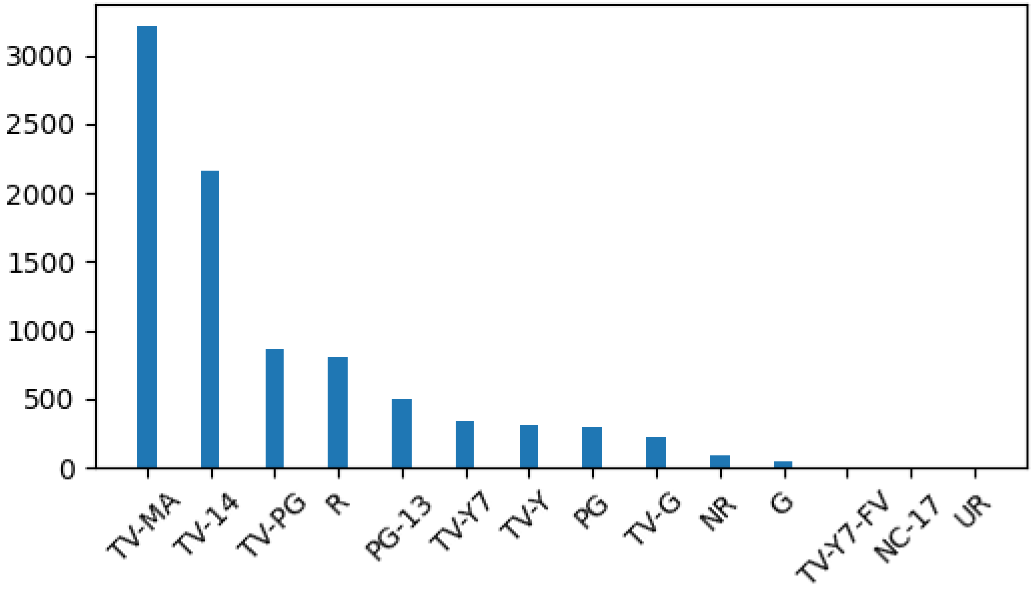
In [1356]:

```
# 4. Distribution of contents added over months on Netflix Platform

plt.figure(figsize = (10,3))
plt.subplot(1,2,1)
count_month = data["month"].value_counts()
plt.bar(count_month.index,count_month,width = 0.3)
plt.subplot(1,2,2)
sns.kdeplot(data["month"])
plt.show()

#Uniform distribution of added content on Netflix over months can be observed. However, a bit dip
#can be observed in month of feburary
```

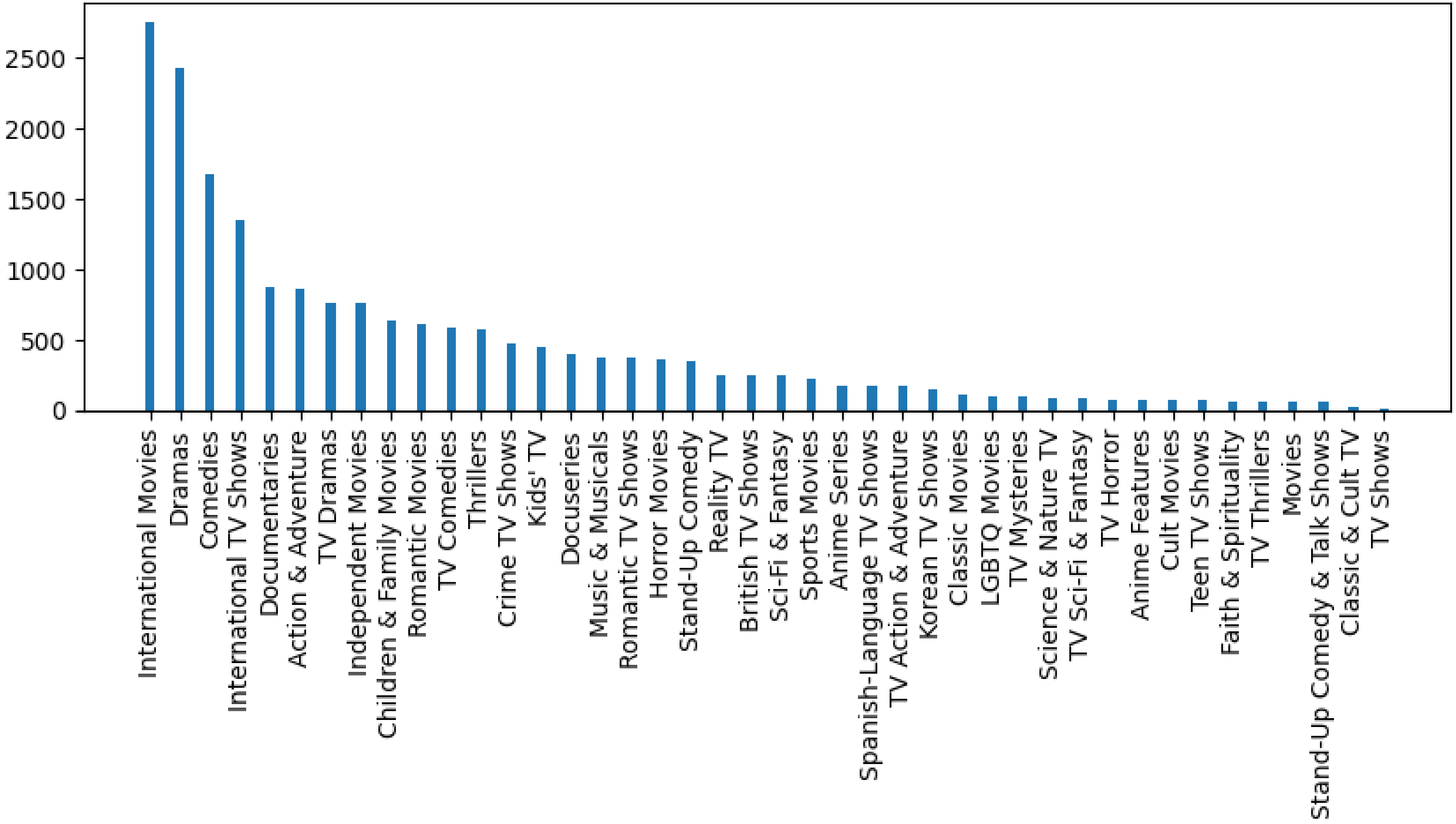


In [1357]:

```
# 5. Category which have been used for rating frequently.

plt.figure(figsize = (6,3))
count_5 = data["rating"].value_counts()
plt.bar(count_5.index,count_5,width = 0.3)
plt.xticks(rotation= 45)
plt.show()

#It can be observed that maximum ratings have been provided under category "TV-MA", followed by "TV-14".
```



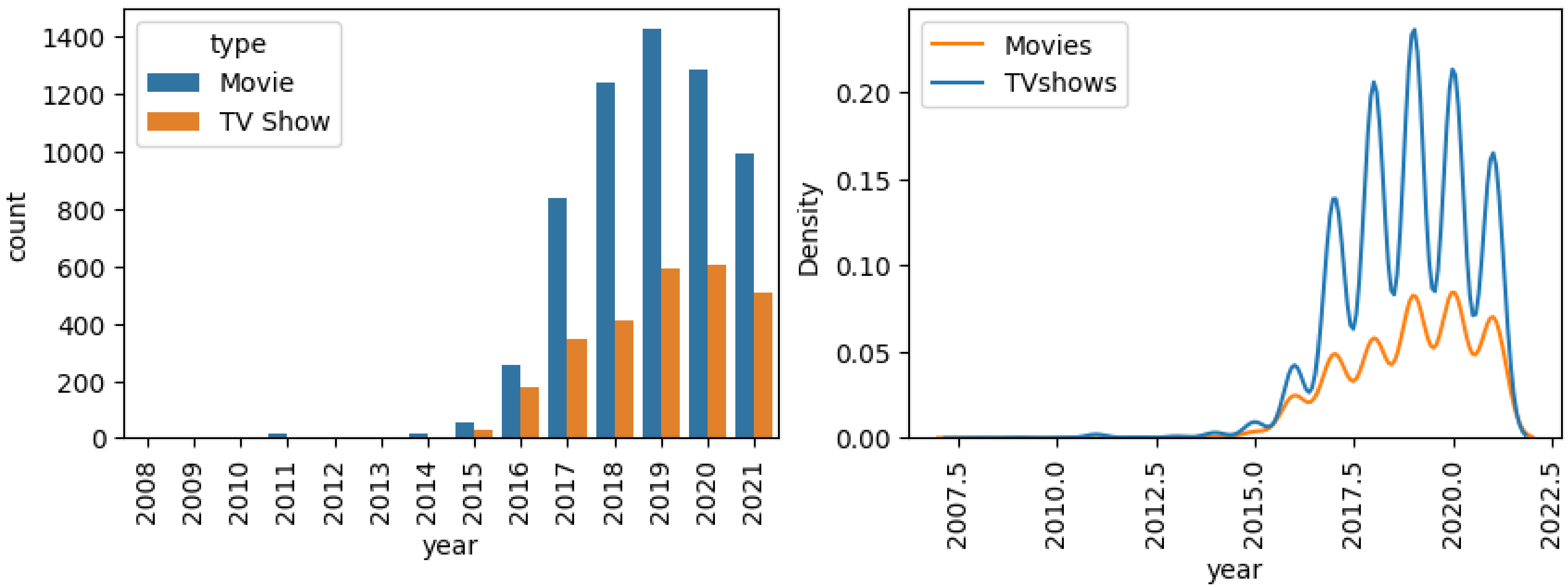
```
In [1358]: # 6. Most popular genres

           plt.figure(figsize = (10,3))
           count_6 = data_genre["genre"].value_counts()
           plt.bar(count_6.index,count_6,width = 0.3)
           plt.xticks(rotation= 90)
           plt.show()

           #People love to watch International Movies followed by the Dramas.
```



VISUAL ANALYSIS: BIVARIATE

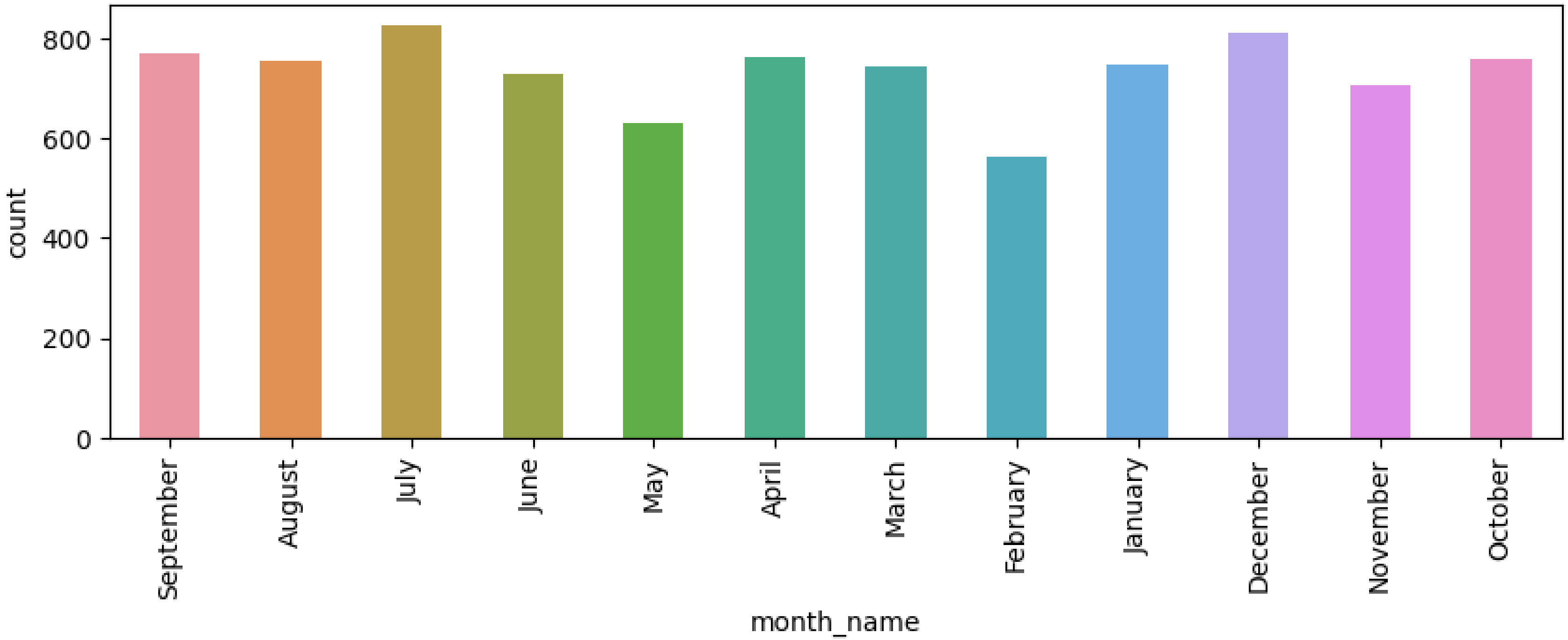
In [1360]:

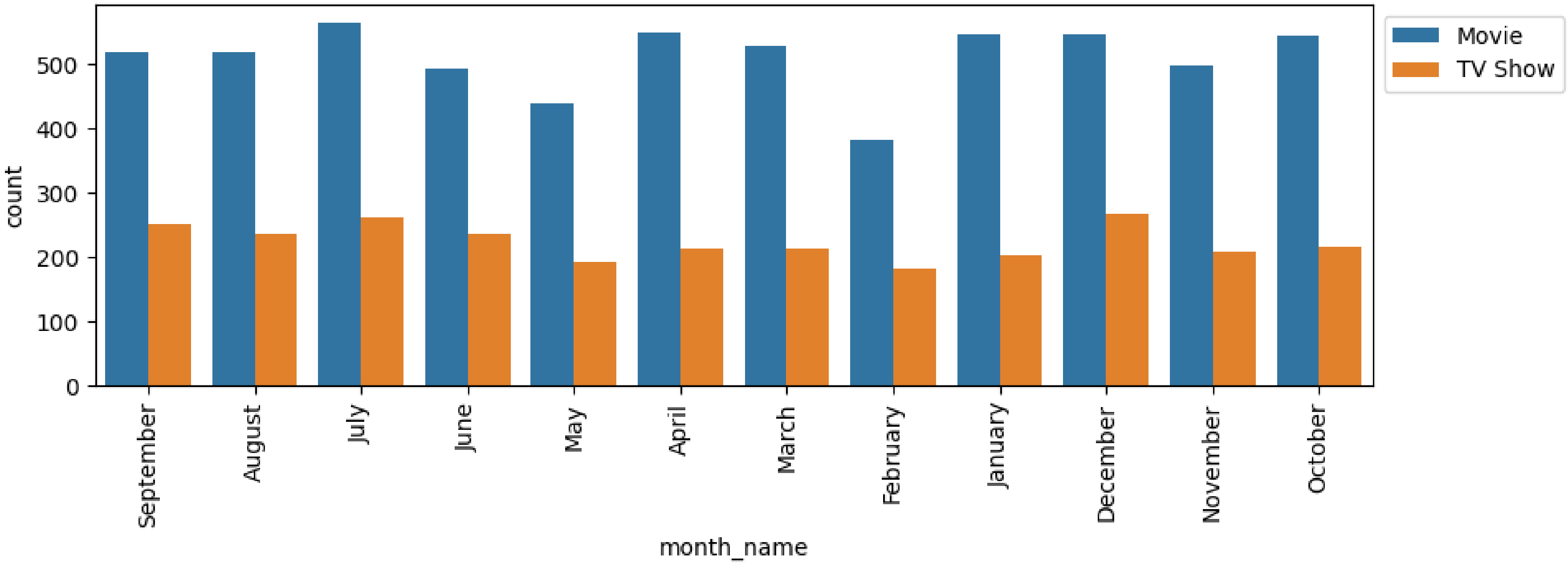
```
# 1. Distribution of type of content added on Netflix Platform over years.

plt.figure(figsize = (10,3))
plt.subplot(1,2,1)
sns.countplot(data = data, x = "year", hue = "type")
plt.xticks(rotation = 90)
plt.subplot(1,2,2)
sns.kdeplot(data = data, x = "year", hue = "type")
plt.xticks(rotation = 90)
plt.legend(["Movies","TVshows"], loc = "upper left")
plt.show()

# Netflix has attained popularity gradually. Netflix is in the market for last 12 years but for the
# last five years, it is attracting customers.
# Netflix has started adding huge content on their platform. In year 2016, the data added was far greater
# than the previous year and from then onwards, the count of added content is increasing.
# Netflix has added their maximum content in year 2019.
# The number of movies added is far greater than TV shows added on their platform. However, from 2019
# to 2021, the Movie/TV Show content been added on their platform has shown a decline trend. But,
# the decline slope of movies is higher than slope of TV Shows.
```


type
Movie
TV Show
count
0
200
400
600
800
1000
1200
1400
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
year
Movies
TVshows
Density
0.00
0.05
0.10
0.15
0.20
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5
year

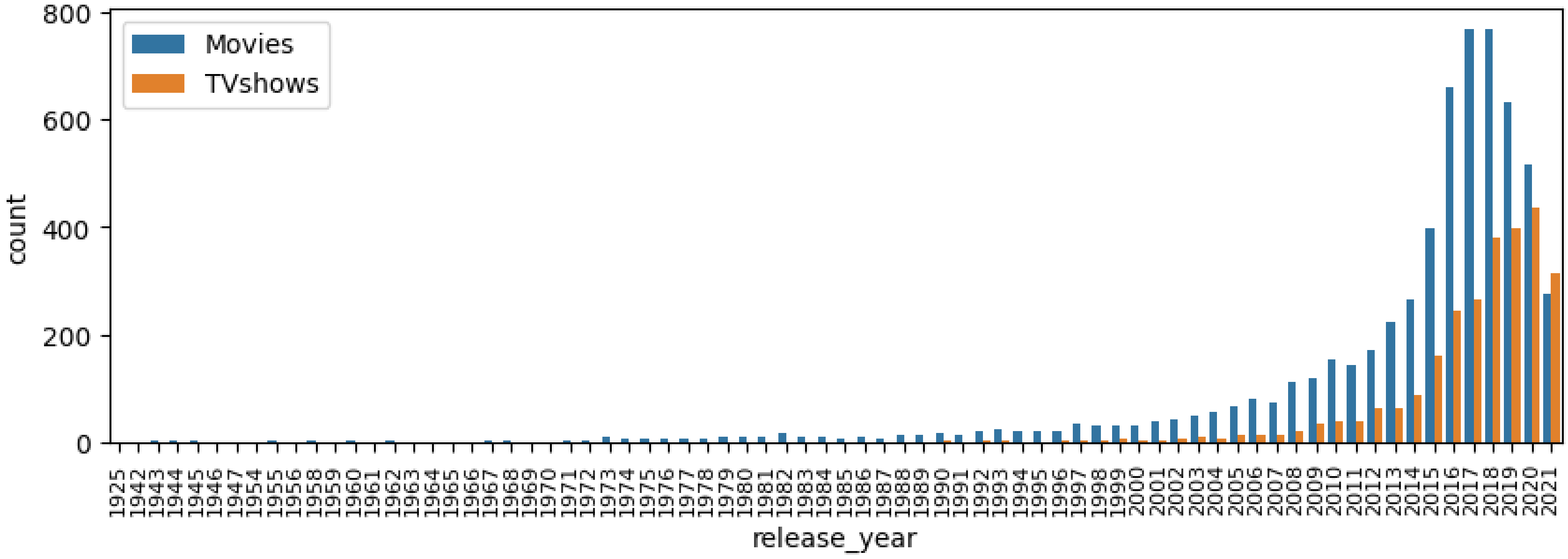
In [1361]:

```
# 2. Distribution of type of content added on Netflix Platform over months.

plt.figure(figsize = (10,3))
plt.xticks(rotation = 90)
sns.countplot(data = data, x = "month_name", width = 0.5)
plt.show()

plt.figure(figsize = (10,3))
plt.xticks(rotation = 90)
sns.countplot(data = data, x = "month_name", hue = "type")
plt.legend(bbox_to_anchor=(1, 1), loc=2)
plt.show()

# According to plot, the feburary and may month has less added data, whereas all other months have considerable
# same amount of added content. The month July and December have maximum added content.
# It can be observed that the Fall(month from July to December) season has large amount of content added compare
# to the remaining months.
```



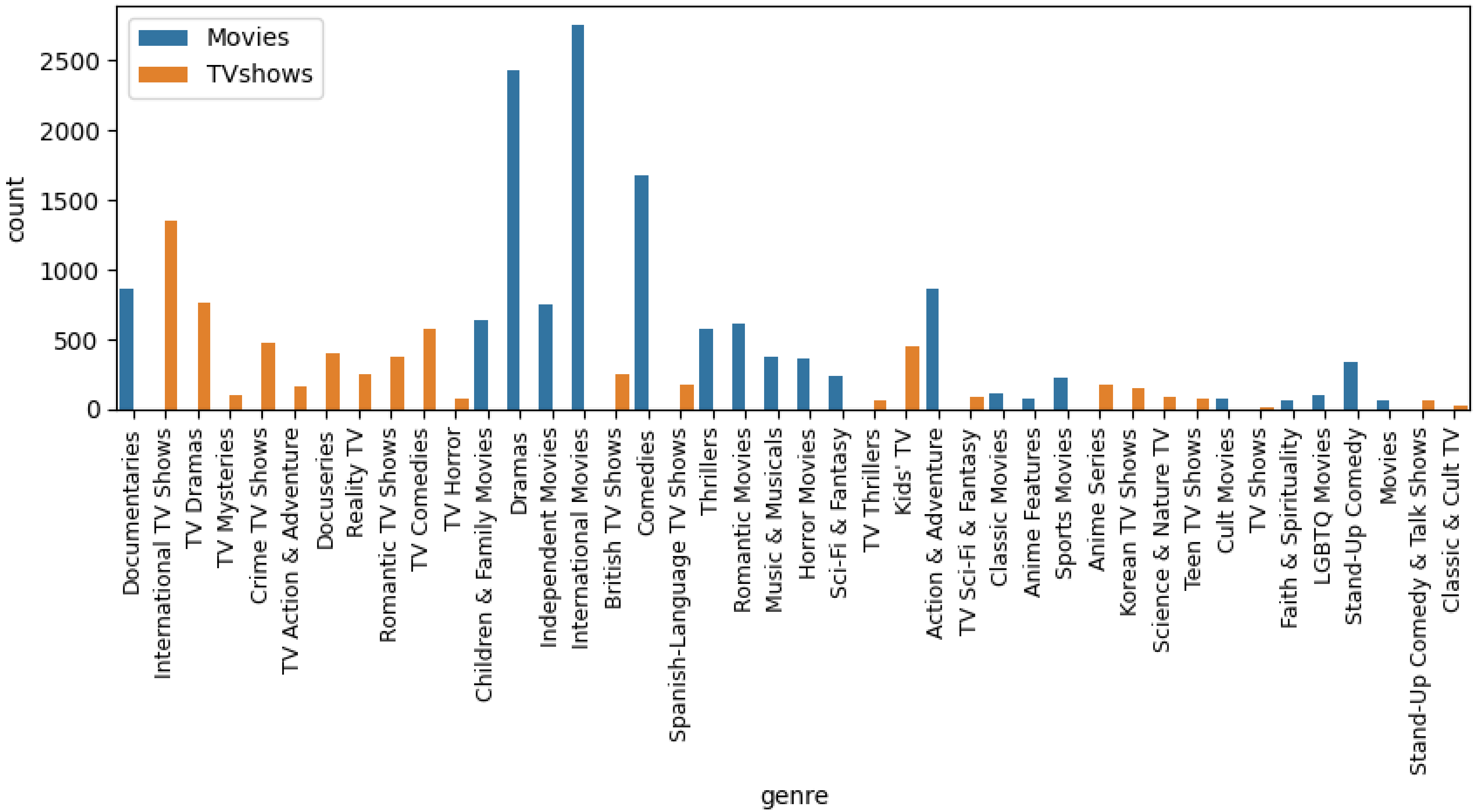

count
0
100
200
300
400
500
September
August
July
June
May
April
March
February
January
December
November
October
month_name
Movie
TV Show

```
In [1364]: # 3. Movies and TVshows release in years

           plt.figure(figsize = (10,3))
           plt.xticks(rotation = 90,fontsize = 8)
           sns.countplot(data = data, x = "release_year", hue = "type")
           plt.legend(["Movies","TVshows"], loc = "upper left")
           plt.show()

           # With time, the interest in watching movies and TVshows has increasing only. However, since 2015, with the
           # easily availability of internet and increasing usage of mobiles, the watching time has been increased
           # abruptly. Earlier, people fond of movies but audience of present time would love to watch not only movies
           # but also TVshows.
           # The creation of good TV shows like "Breaking Bad", "Game of Thrones", "Stranger Things" have diverted the
           # prime time from movies to TV shows.
           # TV shows are so much in demand that in the year 2021, the watching time of TV shows is greater than of movies.
```



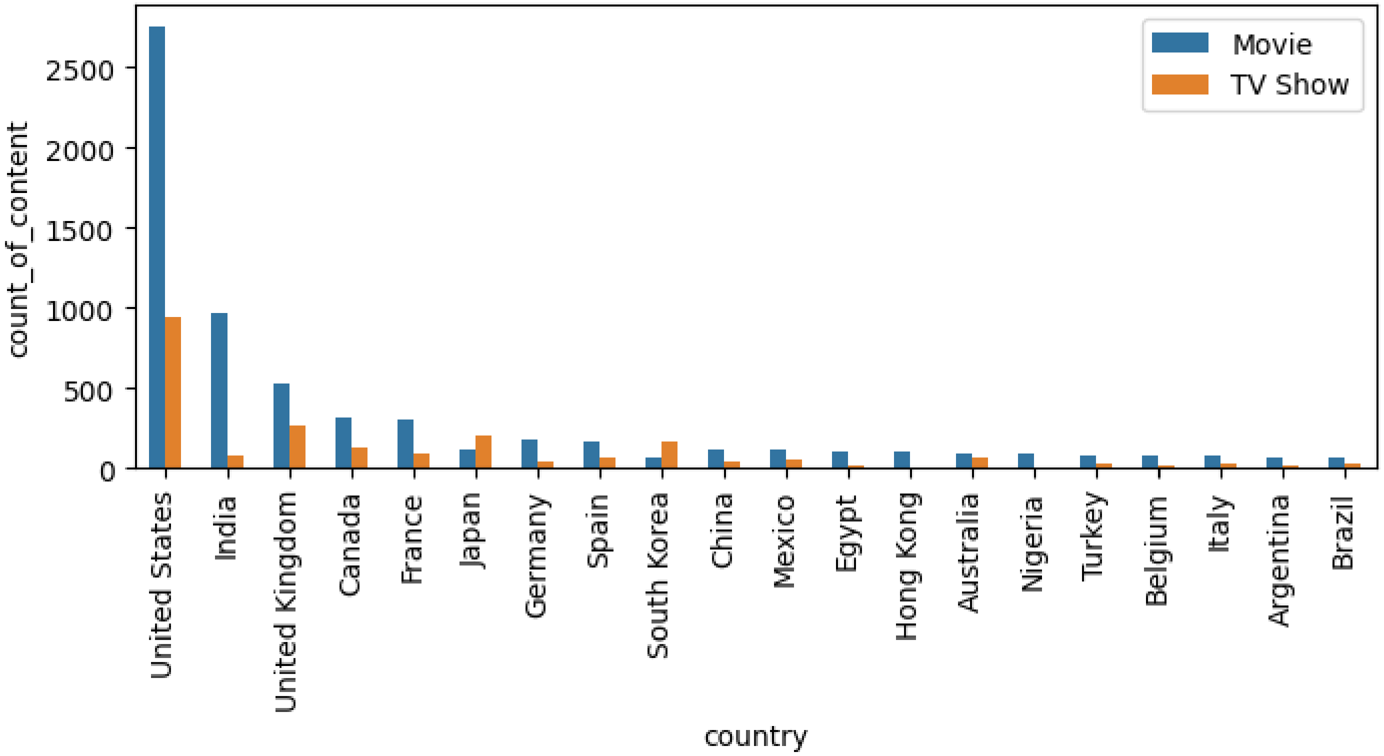
In [1369]:

```
# 4. Variation of Genres over Movies and TV shows.

data_genre_data = data_genre.merge(data, on ="title", how = "left" )
data_genre_data

plt.figure(figsize = (10,3))
plt.xticks(rotation = 90, fontsize = 9)
sns.countplot(data = data_genre_data, x = "genre", hue = "type")
plt.legend(["Movies","TVshows"], loc = "upper left")
plt.show()

# It can be seen from the graph that Movies created are mostly of genres like International Movies, Dramas,
# and Comedies while TV shows are of genres like International TV Shows, TV Dramas, and TV Comedies.
# It can be inferred that people tend to like Content which is Internationally acclaimed, followed by
# Dramas and comedies.
```


Movies
TVshows
count
2500
2000
1500
1000
500
0
genre
Documentaries
International TV Shows
TV Dramas
TV Mysteries
Crime TV Shows
TV Action & Adventure
Docuseries
Reality TV
Romantic TV Shows
TV Comedies
TV Horror
Children & Family Movies
Dramas
Independent Movies
International Movies
British TV Shows
Comedies
Spanish-Language TV Shows
Thrillers
Romantic Movies
Music & Musicals
Horror Movies
Sci-Fi & Fantasy
TV Thrillers
Kids' TV
Action & Adventure
TV Sci-Fi & Fantasy
Classic Movies
Anime Features
Sports Movies
Anime Series
Korean TV Shows
Science & Nature TV
Teen TV Shows
Cult Movies
TV Shows
Faith & Spirituality
LGBTQ Movies
Stand-Up Comedy
Movies
Stand-Up Comedy & Talk Shows
Classic & Cult TV

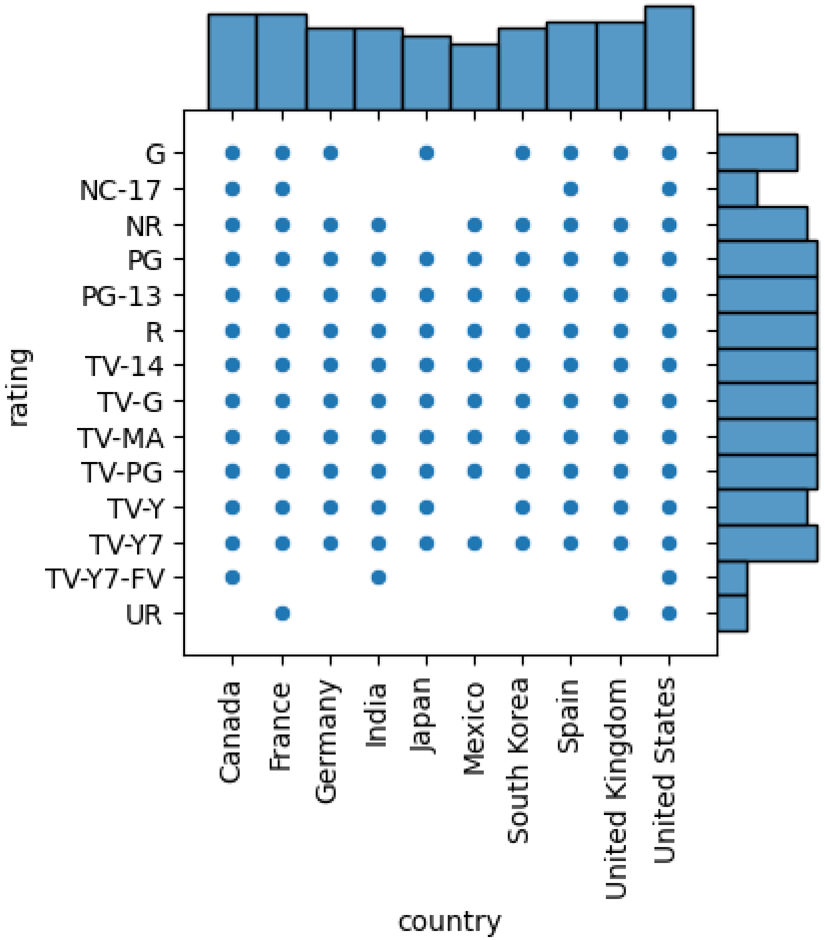
In [1371]:

```
# 5. Movies and TV shows distribution over countries

data_country_data = data_country.merge(data, on ="title", how = "left" )
data_country_data_1 = data_country_data.groupby(["country_x","type"])[["title"]].count().reset_index()
data_country_data_1.rename({"country_x":"country","title":"count_of_content"},axis = 1,inplace = True)
data_country_data_top20 = data_country_data_1.sort_values(by = "count_of_content", ascending = False)
data_country_data_top = data_country_data_top20[data_country_data_top20["country"].isin(top_countries["country"])]

plt.figure(figsize = (8,3))
plt.xticks(rotation = 90)
sns.barplot(data = data_country_data_top, x = "country", y = "count_of_content", hue = "type", width = 0.5)
plt.legend(loc = "upper right")
plt.show()

# The graph has been shown for the top 20 countries having maximum count of Movies/TV Shows.
# People of most of the countries are likely to watch Movies compared to TV Shows. The countries like United
# States and India can be seen having huge difference between viewers of Movies and TV Shows.
# Japan and South Korea are two exceptions in top 20 countries that have more number of viewers in TV Shows
# than Movies.
```

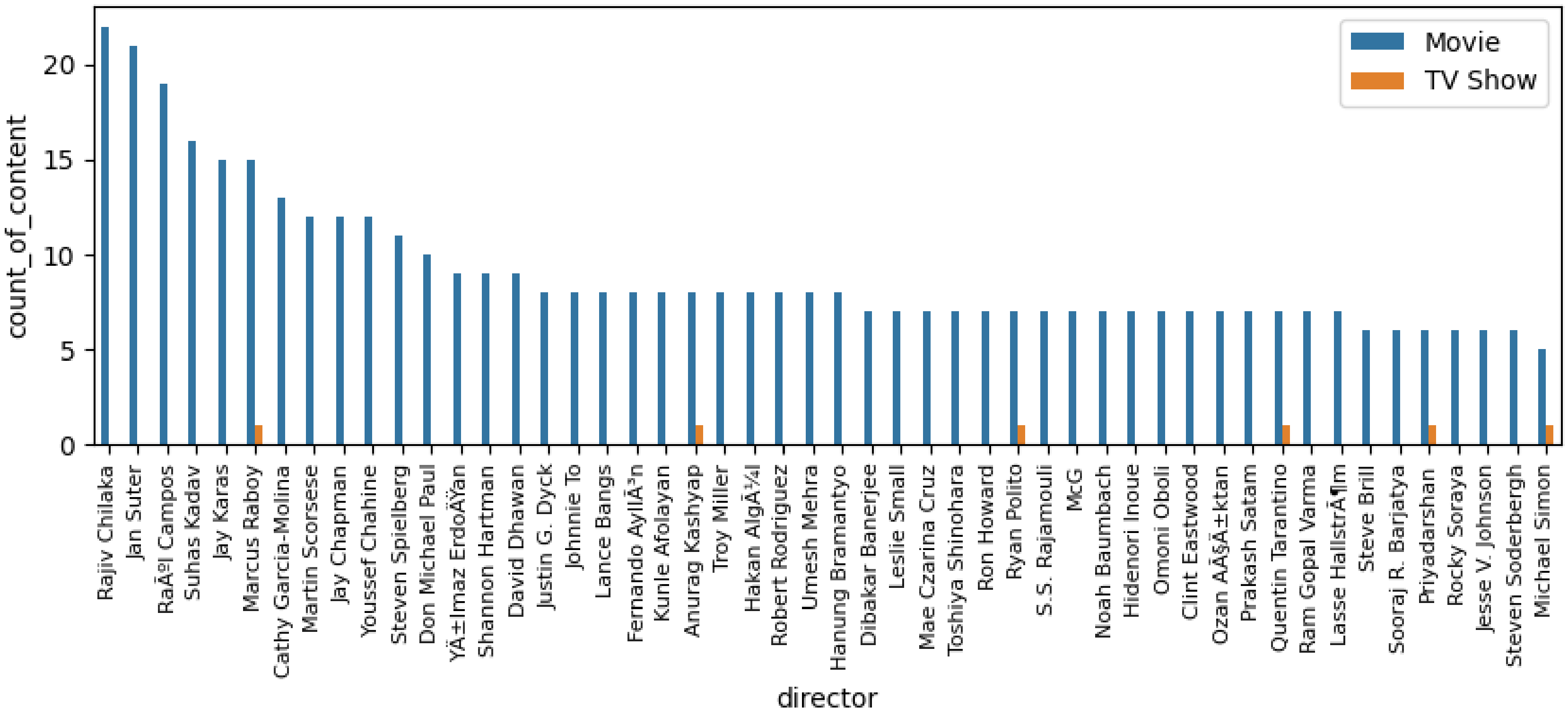

Movie
TV Show
count_of_content
0
500
1000
1500
2000
2500
country
United States
India
United Kingdom
Canada
France
Japan
Germany
Spain
South Korea
China
Mexico
Egypt
Hong Kong
Australia
Nigeria
Turkey
Belgium
Italy
Argentina
Brazil

In [1372]:

```
# 6. Rating distribution over countries

data_country_rating = data_country.merge(data, on ="title", how = "left" )
data_country_rating.drop("country_y",axis =1, inplace = True)
data_country_rating.rename({"country_x":"country"},axis = 1,inplace = True)
data_country_rating_1 = data_country_rating[data_country_rating["country"].isin(top_countries["country"].head(10))]
data_country_rating_2 = data_country_rating_1.groupby(["country","rating"])[["title"]].count().reset_index()
data_country_rating_2.rename({"title":"count_of_content"},axis = 1,inplace = True)

sns.jointplot(data = data_country_rating_2, x = "country", y = "rating",kind='scatter',  height=4, ratio=5, space=0)
plt.xticks(rotation = 90)
plt.show()

#The US content has got all types of ratings, followed by Canada.
#It is more unlikely to get ratings of "UR" and "TV-Y7-FV".
```

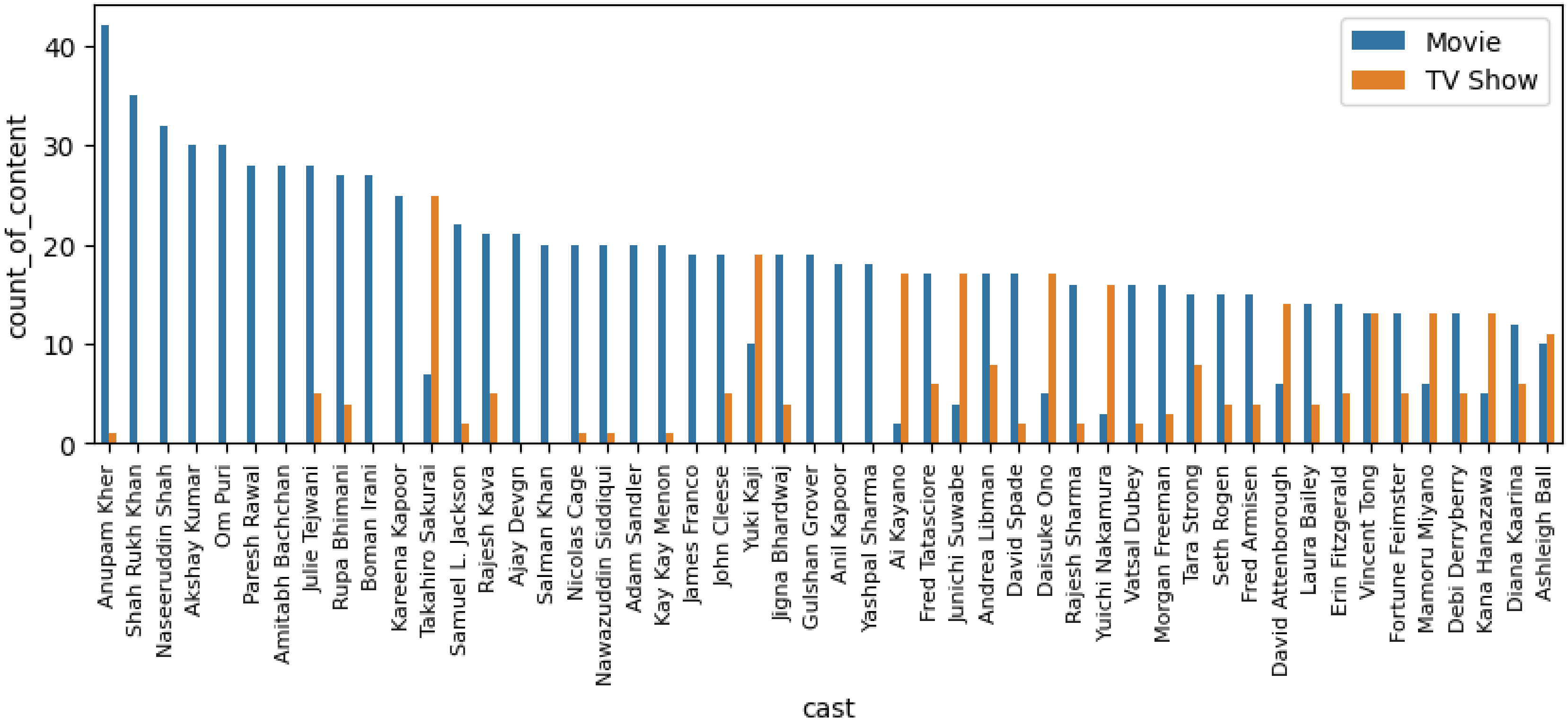

G
NC-17
NR
PG
PG-13
R
TV-14
TV-G
TV-MA
TV-PG
TV-Y
TV-Y7
TV-Y7-FV
UR
rating
Canada
France
Germany
India
Japan
Mexico
South Korea
Spain
United Kingdom
United States
country

```
In [1373]: # 7. Movies and TV shows distribution over Directors

            data_director_data = data_director.merge(data, on ="title", how = "left" )
            data_director_data_1 = data_director_data.groupby(["director_x","type"])[["title"]].count().reset_index()
            data_director_data_1.rename({"director_x":"director","title":"count_of_content"},axis = 1,inplace = True)
            data_director_data_top20 = data_director_data_1.sort_values(by = "count_of_content", ascending = False)
            data_director_data_top = data_director_data_top20[data_director_data_top20["director"].
                                                              isin(top_directors["director"])]
            data_director_data_top
            plt.figure(figsize = (10,3))
            plt.xticks(rotation = 90, fontsize = 8)
            sns.barplot(data = data_director_data_top, x = "director", y = "count_of_content", hue = "type", width = 0.5)
            plt.legend(loc = "upper right")
            plt.show()

            # The graph has been shown for the top 50 directors having maximum count of Movies/TV Shows.
            # It is quite evident that popular directors are more likely to direct movies. There are some exceptions
            # who try their hands on both type of content. "Marcus Raboy", "Anurag Kashyap",  and "Ryan Polito" are
            # such directors who explore area of TV shows also. However, comparison to movies count,number of
            # TV shows directed by them are too less.
```


Movie
TV Show
count_of_content
0
5
10
15
20
Rajiv Chilaka
Jan Suter
RaÃºl Campos
Suhas Kadav
Jay Karas
Marcus Raboy
Cathy Garcia-Molina
Martin Scorsese
Jay Chapman
Youssef Chahine
Steven Spielberg
Don Michael Paul
YÄ±lmaz ErdoÄŸan
Shannon Hartman
David Dhawan
Justin G. Dyck
Johnnie To
Lance Bangs
Fernando AyllÃ³n
Kunle Afolayan
Anurag Kashyap
Troy Miller
Hakan AlgÃ¼l
Robert Rodriguez
Umesh Mehra
Hanung Bramantyo
Dibakar Banerjee
Leslie Small
Mae Czarina Cruz
Toshiya Shinohara
Ron Howard
Ryan Polito
S.S. Rajamouli
McG
Noah Baumbach
Hidenori Inoue
Omoni Oboli
Clint Eastwood
Ozan AÃ§Ä±ktan
Prakash Satam
Quentin Tarantino
Ram Gopal Varma
Lasse HallstrÃ¶m
Steve Brill
Sooraj R. Barjatya
Priyadarshan
Rocky Soraya
Jesse V. Johnson
Steven Soderbergh
Michael Simon
director

In [1375]:

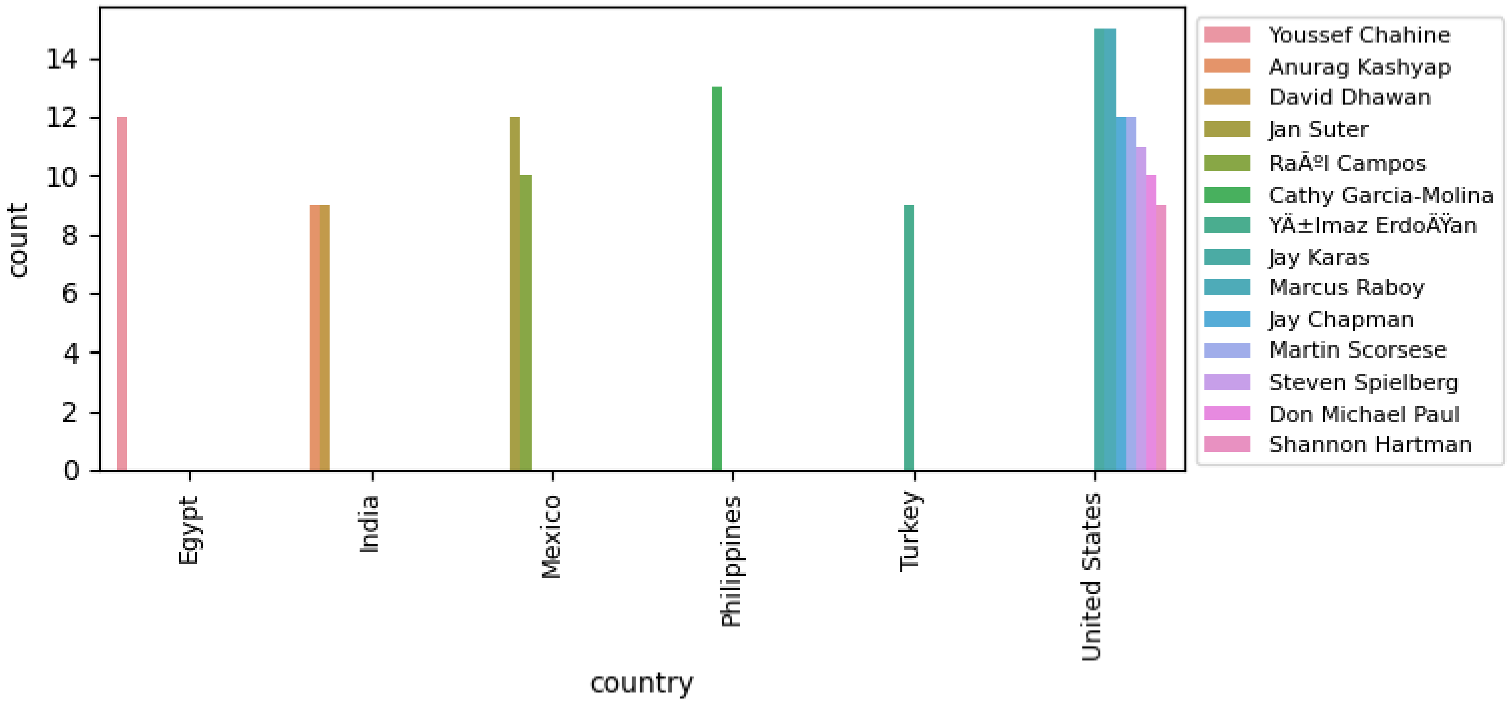
```
# 8. Movies and TV shows distribution over Actors

data_cast_data = data_cast.merge(data, on ="title", how = "left" )
data_cast_data_1 = data_cast_data.groupby(["cast_x","type"])[["title"]].count().reset_index()
data_cast_data_1.rename({"cast_x":"cast","title":"count_of_content"},axis = 1,inplace = True)
data_cast_data_top20 = data_cast_data_1.sort_values(by = "count_of_content", ascending = False)
data_cast_data_top = data_cast_data_top20[data_cast_data_top20["cast"].isin(top_cast["cast"])]
data_cast_data_top
plt.figure(figsize = (10,3))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_cast_data_top, x = "cast", y = "count_of_content", hue = "type", width = 0.5)
plt.legend(loc = "upper right")
plt.show()

# The graph has been shown for the top 50 actors having maximum count of Movies/TV Shows.
# The most popular actors are more likely to act in movies instead of TV Shows. The most popular actor
# "Anupam Kher" has acted in both movies and shows. Similarly, from top10 popular actors, only three
# actors(Anupam Kher, Julie Tejwani and Rupa Bhimani) have acted in both types of content.
# There are certain actors who have acted in TV shows and gain equal popularity. Takahiro Sakurai,
# Yuki Kaji, Ai Kayano, and others are such actors. It's important to note here that these actors
# are from Japan where TV shows are more popular than movies.
#Actors who are among famous top25 to top50 list, are equally devoted to both type of content.
```


count_of_content
0
10
20
30
40
Movie
TV Show
cast
Anupam Kher
Shah Rukh Khan
Naseeruddin Shah
Akshay Kumar
Om Puri
Paresh Rawal
Amitabh Bachchan
Julie Tejwani
Rupa Bhimani
Boman Irani
Kareena Kapoor
Takahiro Sakurai
Samuel L. Jackson
Rajesh Kava
Ajay Devgn
Salman Khan
Nicolas Cage
Nawazuddin Siddiqui
Adam Sandler
Kay Kay Menon
James Franco
John Cleese
Yuki Kaji
Jigna Bhardwaj
Gulshan Grover
Anil Kapoor
Yashpal Sharma
Ai Kayano
Fred Tatasciore
Junichi Suwabe
Andrea Libman
David Spade
Daisuke Ono
Rajesh Sharma
Yuichi Nakamura
Vatsal Dubey
Morgan Freeman
Tara Strong
Seth Rogen
Fred Armisen
David Attenborough
Laura Bailey
Erin Fitzgerald
Vincent Tong
Fortune Feimster
Mamoru Miyano
Debi Derryberry
Kana Hanazawa
Diana Kaarina
Ashleigh Ball

In [1376]:
```
# 9. Talent of direction content from various countries

data_country_director = data_country.merge(data_director, on = "title")
data_country_director1 = data_country_director.groupby(["country","director"])[["title"]].count()
data_country_director1.rename({"title":"count"}, axis = 1, inplace = True)

data_country_director2 =data_country_director1[data_country_director1["count"] > 5].reset_index()

data_country_director3 = data_country_director2.sort_values(by =["country","count"],ascending =[True, False])

#Lets observe those countries with directors who at least directed 9 movies/TV shows

data_country_director4 = data_country_director3[data_country_director3["count"] >= 9]
display(data_country_director4.head())
print("============================================================================================")

plt.figure(figsize = (7,3))
plt.xticks(rotation = 90, fontsize = 9)
sns.barplot(data = data_country_director4, x = "country", y = "count", hue = "director")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()

# The United States have most directors who have directed more than 9 movies.
# Jay Karas and Marcus Raboy directed movies from United States have been attained the maximum place in
# Netflix platform.
# Similarly, Youssef Chahine from Egypt and Jan Suter from Mexico have directed more than 10 movies/shows
# and have been able to grab place in Netflix.
# Most Indian movies/TV Shows on Netflix platform have been directed by Anurag Kashyap and David Dhawan.
```

	country	director	count
2	Egypt	Youssef Chahine	12
6	India	Anurag Kashyap	9
8	India	David Dhawan	9
22	Mexico	Jan Suter	12
23	Mexico	RaÃºl Campos	10

```
============================================================================================
```

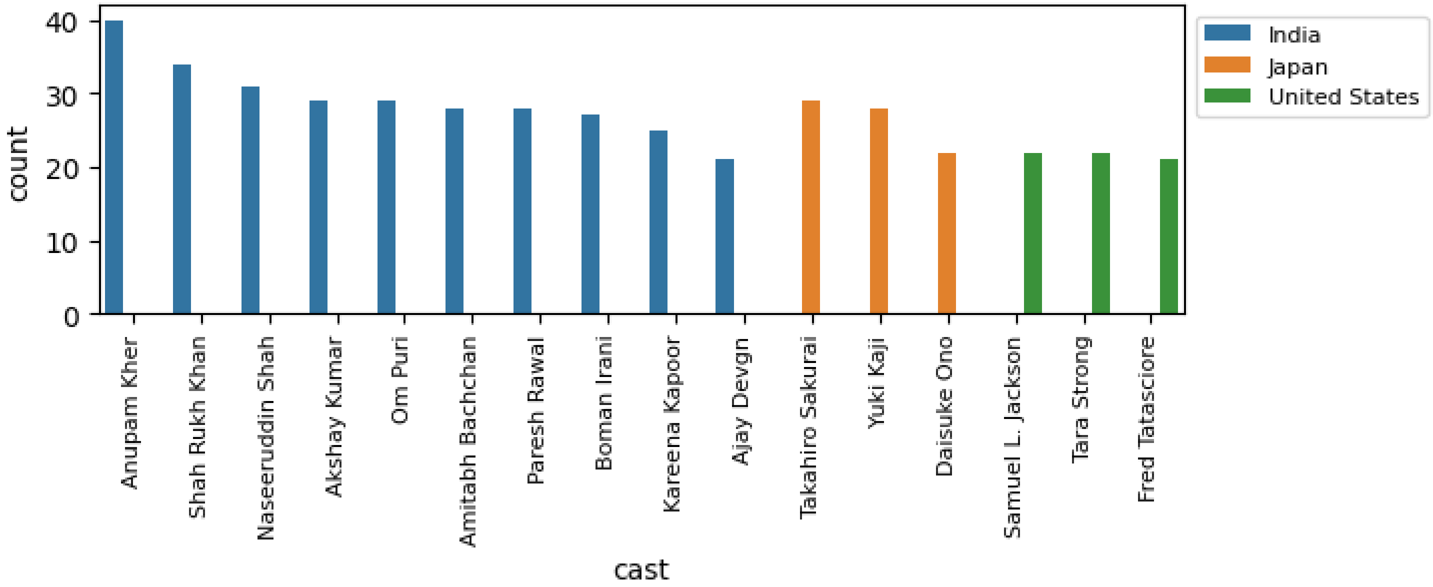

count
0
2
4
6
8
10
12
14
country
Egypt
India
Mexico
Philippines
Turkey
United States
Youssef Chahine
Anurag Kashyap
David Dhawan
Jan Suter
RaÃºl Campos
Cathy Garcia-Molina
YÄ±lmaz ErdoÄŸan
Jay Karas
Marcus Raboy
Jay Chapman
Martin Scorsese
Steven Spielberg
Don Michael Paul
Shannon Hartman

In [1377]:

```
# 10. Talent of Acting from various countries

data_country_cast = data_country.merge(data_cast, on = "title")
data_country_cast1 = data_country_cast.groupby(["country","cast"])[["title"]].count()
data_country_cast1.rename({"title":"count"}, axis = 1, inplace = True)

data_country_cast2 =data_country_cast1[data_country_cast1["count"] > 20].reset_index()

data_country_cast3 = data_country_cast2.sort_values(by = ["country","count"], ascending = [True, False])

display(data_country_cast3.head())
print("============================================================================================")

plt.figure(figsize = (7,2))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_country_cast3, x = "cast", y = "count", hue = "country")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()

#Indian Actors have acted in most of the added movies or shows of Netflix.
#Anupam Kher from India has acted in 40 movies/shows followed by Shahrukh Khan. However,Shahrukh khan has only
#acted in movies (concluded in above graphs).
#Takahiro Sakurai from Japan has acted in about 29 TV shows.
#Samuel L. Jackson from United States has acted in 22 movies/TV shows.
```

	country	cast	count
3	India	Anupam Kher	40
9	India	Shah Rukh Khan	34
6	India	Naseeruddin Shah	31
1	India	Akshay Kumar	29
7	India	Om Puri	29

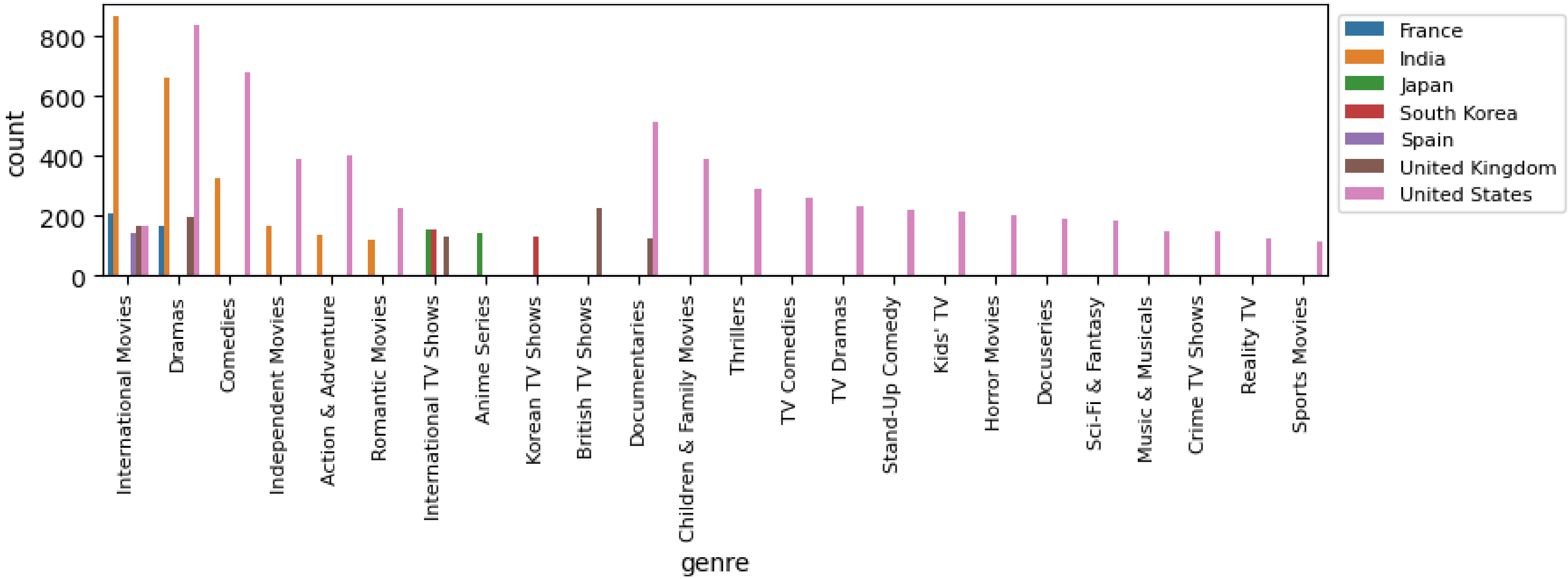
```
============================================================================================
```


count
40
30
20
10
0
cast
Anupam Kher
Shah Rukh Khan
Naseeruddin Shah
Akshay Kumar
Om Puri
Amitabh Bachchan
Paresh Rawal
Boman Irani
Kareena Kapoor
Ajay Devgn
Takahiro Sakurai
Yuki Kaji
Daisuke Ono
Samuel L. Jackson
Tara Strong
Fred Tatasciore
India
Japan
United States

```
In [1249]: # 11. Popular genres in various countries

           data_country_genre = data_country.merge(data_genre, on = "title")
           data_country_genre1 = data_country_genre.groupby(["country","genre"])[["title"]].count()
           data_country_genre1.rename({"title":"count"}, axis = 1, inplace = True)

           data_country_genre2 =data_country_genre1[data_country_genre1["count"] > 100].reset_index()

           data_country_genre3 = data_country_genre2.sort_values(by = ["country","count"], ascending = [True, False])

           display(data_country_genre3.head())
           print("============================================================================================")

           plt.figure(figsize = (9,2))
           plt.xticks(rotation = 90, fontsize = 8)
           sns.barplot(data = data_country_genre3, x = "genre", y = "count", hue = "country")
           plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
           plt.show()

           # It can be observed that United States has a very versatile movie/shows production, followed by India
           # and United Kingdom.
           # United States have movies with maximum kind of genres.However, the most popular genre among US people
           # is Dramas and Comedies.
           # Some people prefer documentaries also in US.
           # Indian people like movies of genre International Movies,Dramas, and Comedies. However, some selective
           # Indian audiences prefer action and romantic movies also.
           # People of France prefers only International Movies and Dramas.
           # Japanese are much into TV Shows. The most popular genres in Japan are International TV Shows and
           # Anime Series.
           # Similar to Japan, South Koreans also prefer TV Shows. The most popular genres in South Korea are
           # International TV Shows and Korean TV shows. The people are so interested in TV shows that the whole
           # category of genre has been assigned to it.
```

	country	genre	count
1	France	International Movies	207
0	France	Dramas	167
6	India	International Movies	864
4	India	Dramas	662
3	India	Comedies	323

```
====================================================================================================
```



In [1378]:

```
# 12. Most popular Director-Actor Pair

data_director_cast = data_director.merge(data_cast, on = "title")
data_director_cast1 = data_director_cast.groupby(["director","cast"])[["title"]].count()
data_director_cast1.rename({"title":"count"}, axis = 1, inplace = True)

data_director_cast2 =data_director_cast1[data_director_cast1["count"] > 5].reset_index()

data_director_cast3 = data_director_cast2.sort_values(by = ["director","count"], ascending = [True, False])

display(data_director_cast3.head())

#Some best Director-Actor pairs are:
#The Cathy Garcia-Molina and Cathy Garcia-Molina of Phillipines.
#David Dhawan and Anupam Kher from India
#Fernando AyllÃ³n and Ricardo Quevedo from Spain and Colombia

#Exceptions:
#S.S. Rajamouli had directed "Bahubali-The Beginning" and "Bahubali-The conclusion", but as this movie has
#been released in multiple languages, the count has been increased.
#Rajiv Chilaka had many casts on "Chota-bheem" movie series.
#Toshiya Shinohara from Japan has directed the series of Movie "InuYasha" which has many casts.
```

	director	cast	count
0	Cathy Garcia-Molina	Joross Gamboa	6
1	David Dhawan	Anupam Kher	6
2	Fernando AyllÃ³n	Ricardo Quevedo	6
3	Hakan AlgÃ¼l	Ata Demirer	6
4	Joey So	Joseph May	6

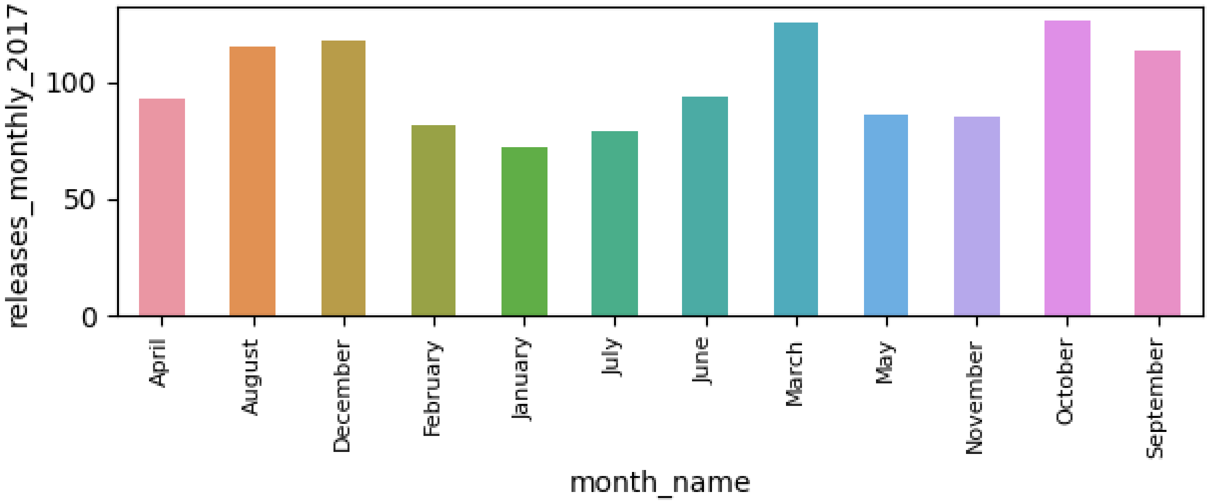
```
In [1379]: # 12. Alternatively Most popular Actor-Director Pair

           data_director_cast = data_director.merge(data_cast, on = "title")
           data_director_cast1 = data_director_cast.groupby(["cast","director"])[["title"]].count()
           data_director_cast1.rename({"title":"count"}, axis = 1, inplace = True)

           data_director_cast2 =data_director_cast1[data_director_cast1["count"] > 5].reset_index()

           data_director_cast3 = data_director_cast2.sort_values(by = ["cast","count"], ascending = [True, False])

           display(data_director_cast3.head())
```

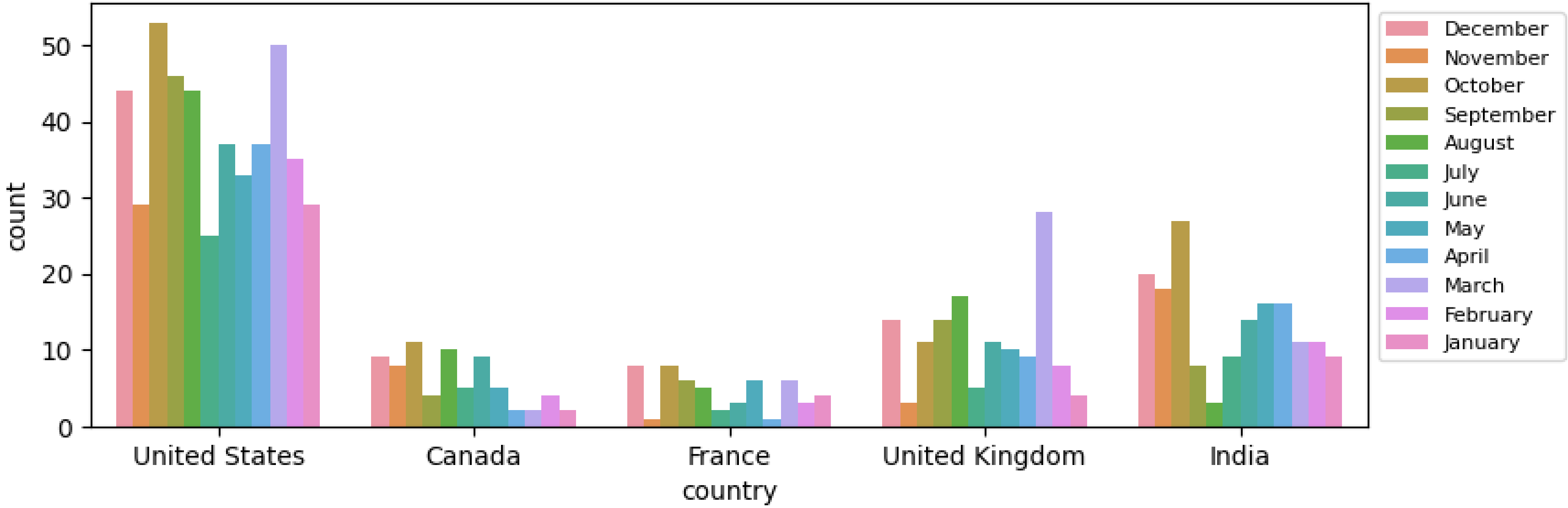
	cast	director	count
0	Anupam Kher	David Dhawan	6
1	Anushka Shetty	S.S. Rajamouli	7
2	Ata Demirer	Hakan AlgÃ¼l	6
3	Houko Kuwashima	Toshiya Shinohara	7
4	Jigna Bhardwaj	Rajiv Chilaka	18

YEAR AND COUNTRY WISE ANALYSIS

In [1380]:

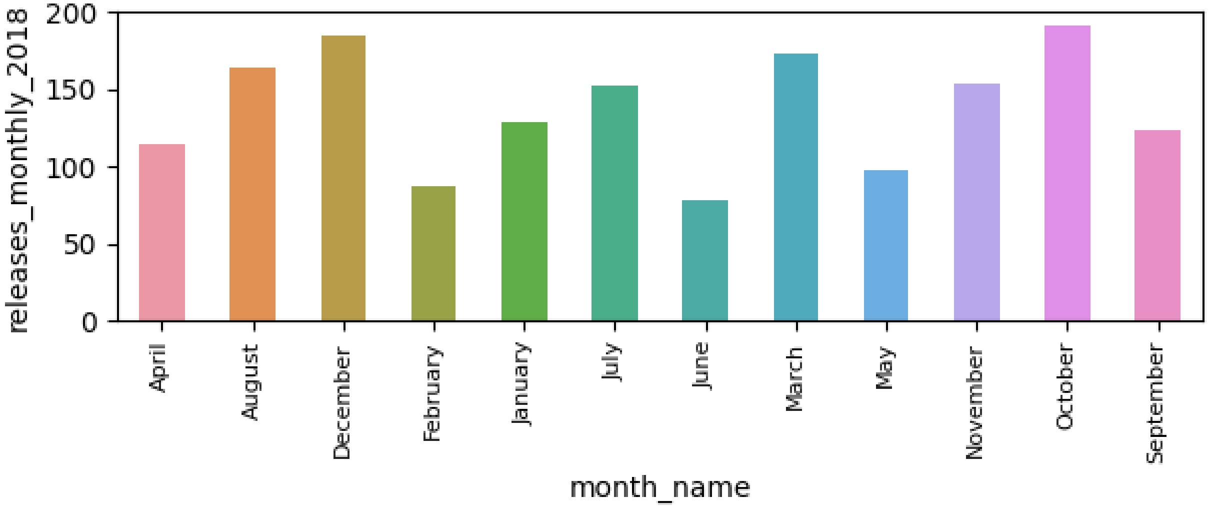
```
#1_a. Year 2017

data_2017 = data[data["year"] == 2017]
data_2017_monthly = data_2017.groupby("month_name")[["title"]].count().reset_index()
data_2017_monthly.rename({"title":"releases_monthly_2017"},axis = 1, inplace = True)
data_2017_monthly

plt.figure(figsize = (7,2))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_2017_monthly, x = "month_name", y = "releases_monthly_2017", width = 0.5)
plt.show()

#1_b Year 2017 and top countries

data2017 = data[["type","title","country","year","month_name"]]
data_2017_country = data_country.merge(data2017, on = "title", how = "left")
data_2017_country.drop("country_y",axis = 1,inplace = True)
data_2017_country.rename({"country_x":"country"},axis =1, inplace = True)
data_2017_country_1 = data_2017_country[(data_2017_country["year"] == 2017)
                                        & (data_2017_country["country"].isin(top_countries["country"].head()))]

plt.figure(figsize = (9,3))
sns.countplot(data = data_2017_country_1, x = "country", hue = "month_name")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()


#In 2017, March and October are the months where highest number of content has been added in Netflix platform.
#With respect to top countries, maximum data has been added in march month in US and UK.
#maximum data has been added in October in US and India.
```

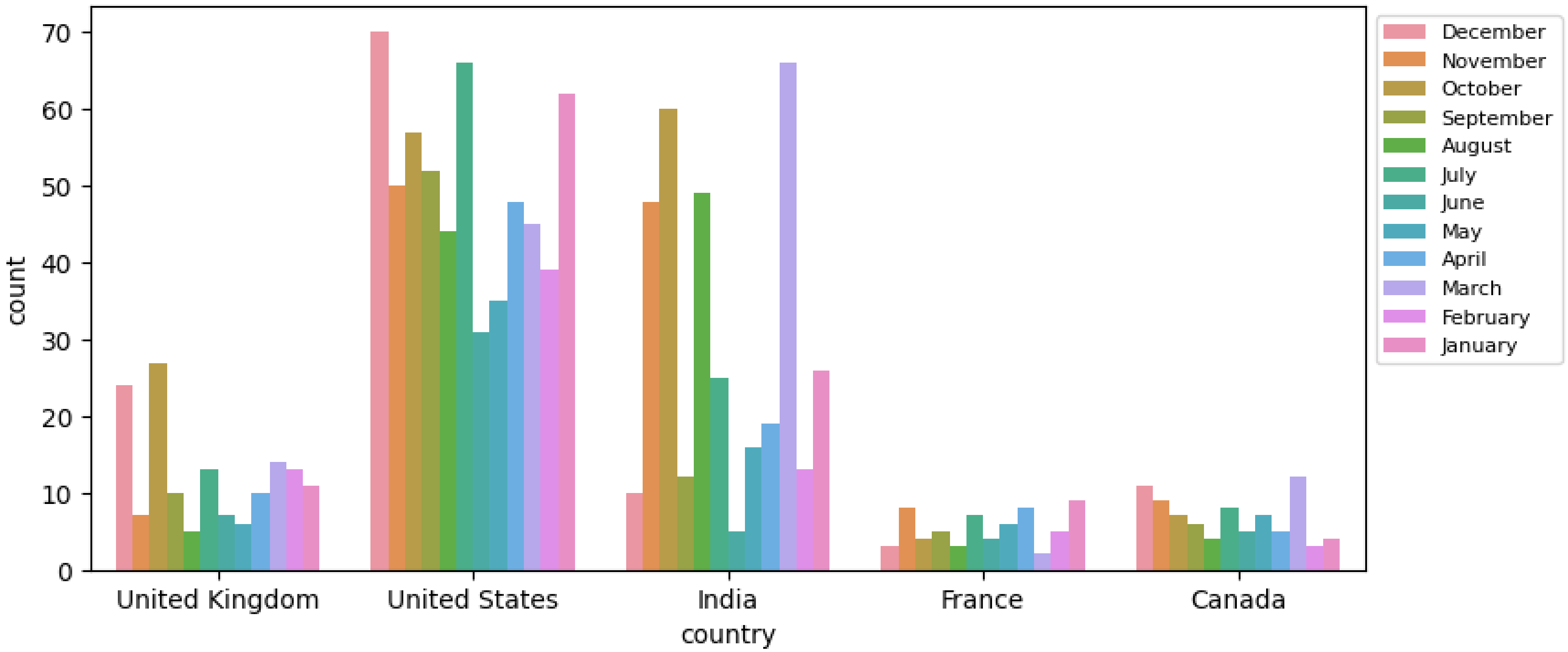

releases_monthly_2017
0
50
100
April
August
December
February
January
July
June
March
May
November
October
September
month_name


count
0
10
20
30
40
50
United States
Canada
France
United Kingdom
India
country
December
November
October
September
August
July
June
May
April
March
February
January

In [1381]:

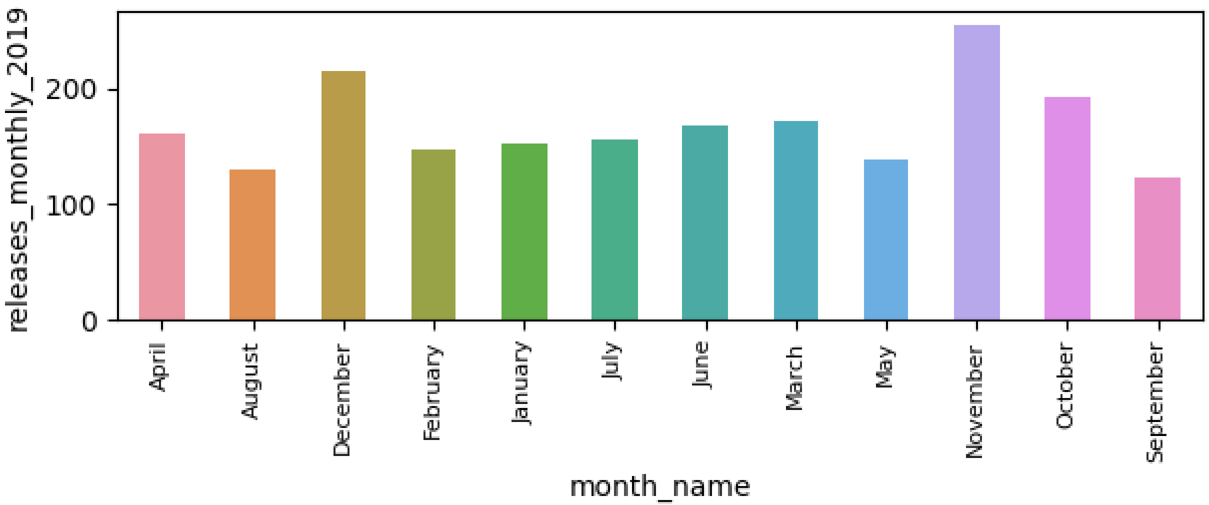
```
#2_a. Year 2018

data_2018 = data[data["year"] == 2018]
data_2018_monthly = data_2018.groupby("month_name")[["title"]].count().reset_index()
data_2018_monthly.rename({"title":"releases_monthly_2018"},axis = 1, inplace = True)
data_2018_monthly

plt.figure(figsize = (7,2))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_2018_monthly, x = "month_name", y = "releases_monthly_2018", width = 0.5)
plt.show()

#2_b Year 2018 and top countries

data2018 = data[["type","title","country","year","month_name"]]
data_2018_country = data_country.merge(data2018, on = "title", how = "left")
data_2018_country.drop("country_y",axis = 1,inplace = True)
data_2018_country.rename({"country_x":"country"},axis =1, inplace = True)
data_2018_country_1 = data_2018_country[(data_2018_country["year"] == 2018)
                                        & (data_2018_country["country"].isin(top_countries["country"].head()))]

plt.figure(figsize = (9,4))
sns.countplot(data = data_2018_country_1, x = "country", hue = "month_name")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()

#In 2018,content added throughout year.However, there is a drop in Feburary,May and June Month.
#With respect to countries, US content has been added throughout year. Similar trend can be shown in India also
#but with exception in months of June,September and December.
```

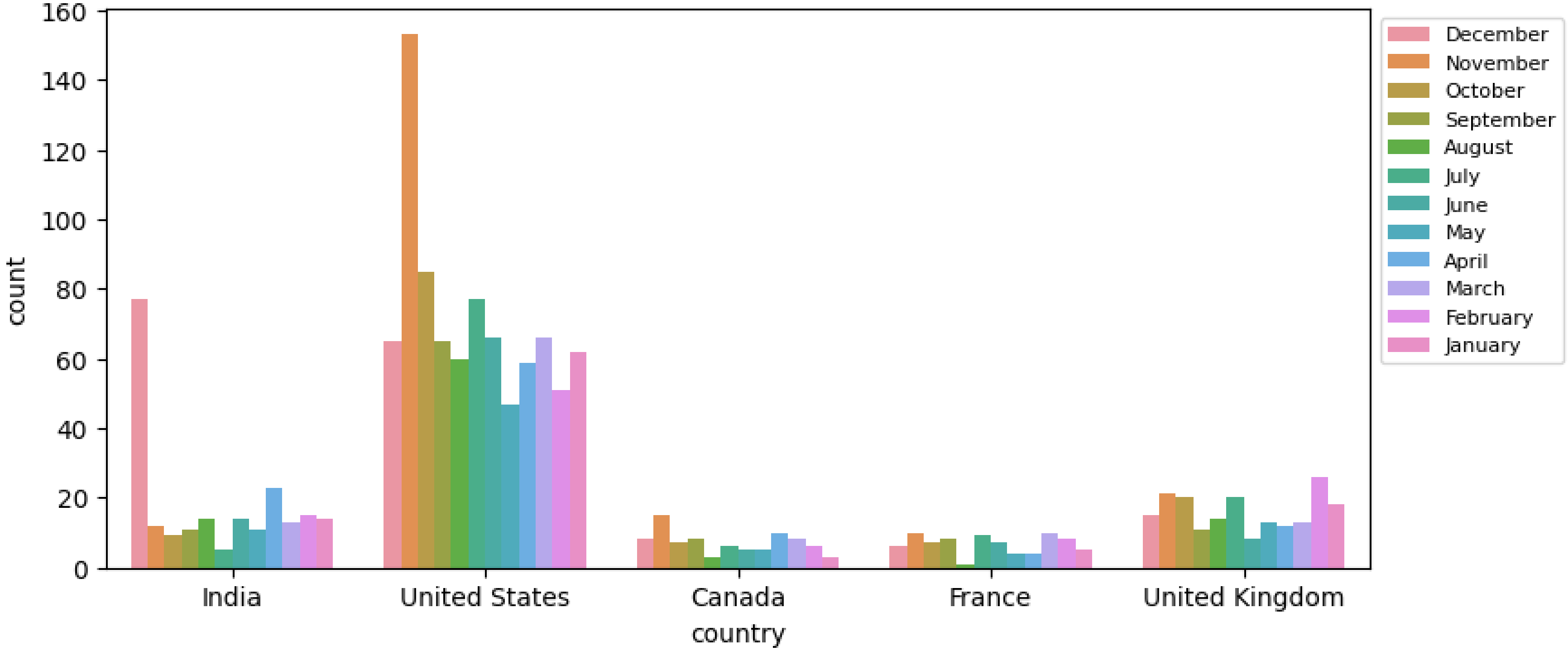

releases_monthly_2018
0
50
100
150
200
April
August
December
February
January
July
June
March
May
November
October
September
month_name


count
0
10
20
30
40
50
60
70
United Kingdom
United States
India
France
Canada
country
December
November
October
September
August
July
June
May
April
March
February
January

In [1382]:

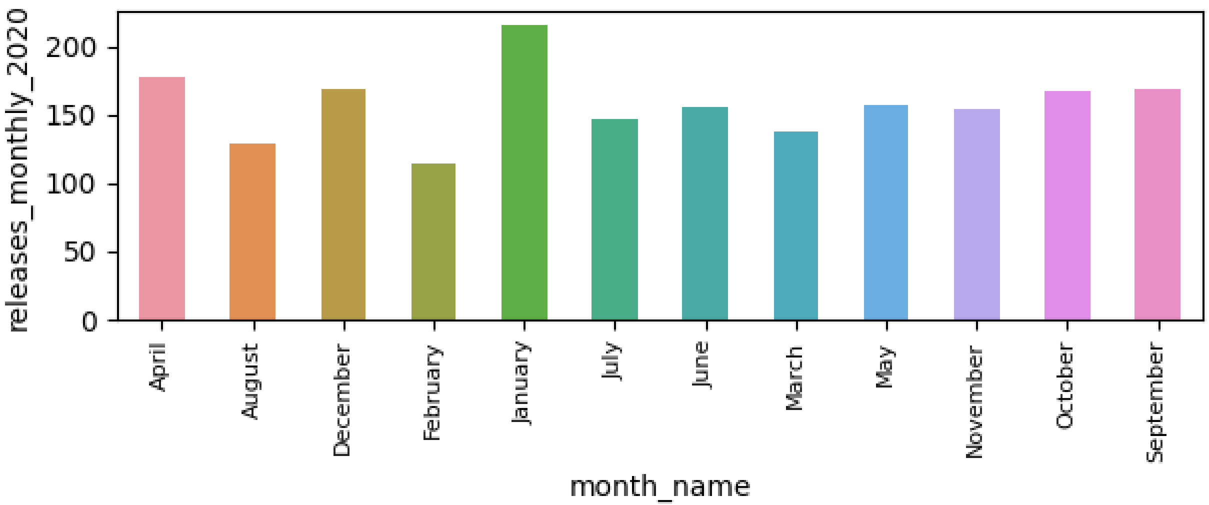
```
#3_a. Year 2019

data_2019 = data[data["year"] == 2019]
data_2019_monthly = data_2019.groupby("month_name")[["title"]].count().reset_index()
data_2019_monthly.rename({"title":"releases_monthly_2019"},axis = 1, inplace = True)
data_2019_monthly

plt.figure(figsize = (7,2))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_2019_monthly, x = "month_name", y = "releases_monthly_2019", width = 0.5)
plt.show()

#3_b Year 2019 and top countries

data2019 = data[["type","title","country","year","month_name"]]
data_2019_country = data_country.merge(data2019, on = "title", how = "left")
data_2019_country.drop("country_y",axis = 1,inplace = True)
data_2019_country.rename({"country_x":"country"},axis =1, inplace = True)
data_2019_country_1 = data_2019_country[(data_2019_country["year"] == 2019)
                                        & (data_2019_country["country"].isin(top_countries["country"].head()))]

plt.figure(figsize = (9,4))
sns.countplot(data = data_2019_country_1, x = "country", hue = "month_name")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()

#In 2019 also,content added throughout year.
#With respect to countries, Most of the US content has been added throughout year.
#In December, a sharp increase of addition of Indian content can be observed.
```

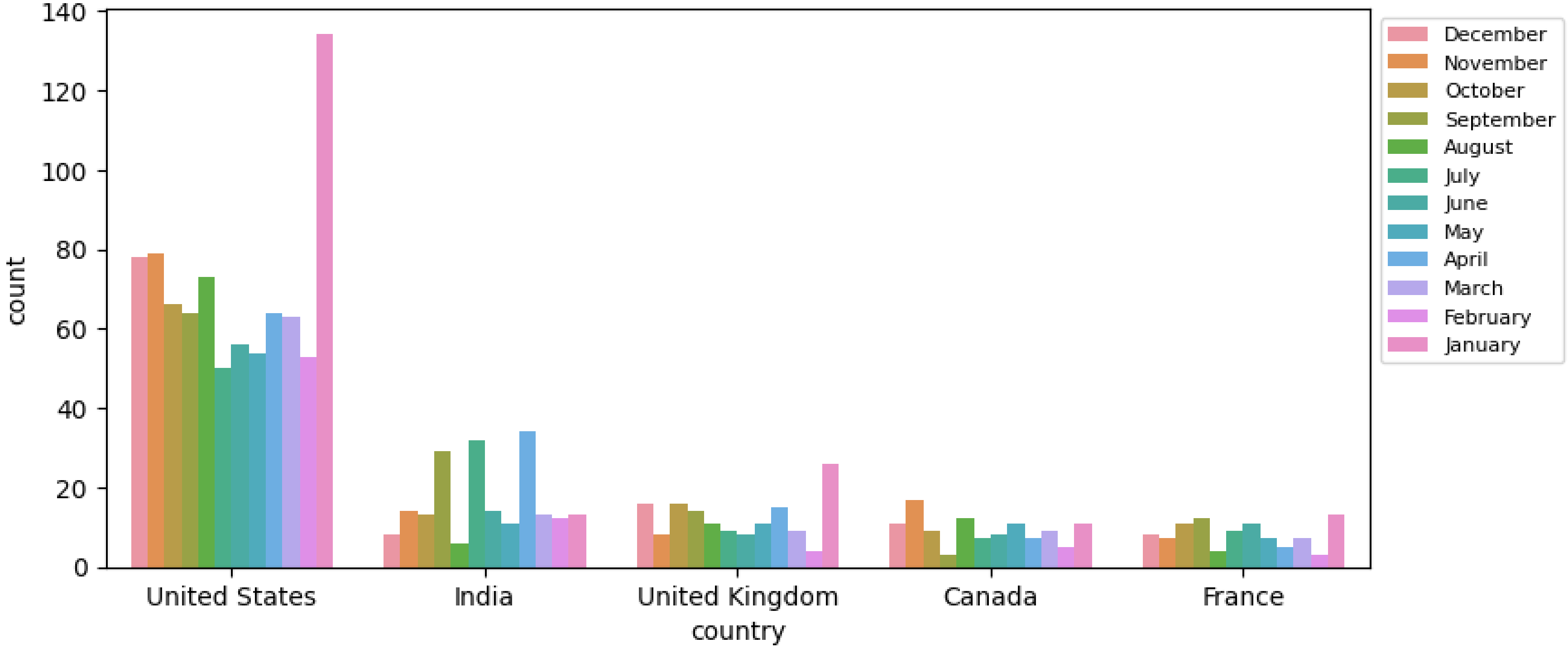

releases_monthly_2019
0
100
200
April
August
December
February
January
July
June
March
May
November
October
September
month_name


count
0
20
40
60
80
100
120
140
160
India
United States
Canada
France
United Kingdom
country
December
November
October
September
August
July
June
May
April
March
February
January

In [1383]:

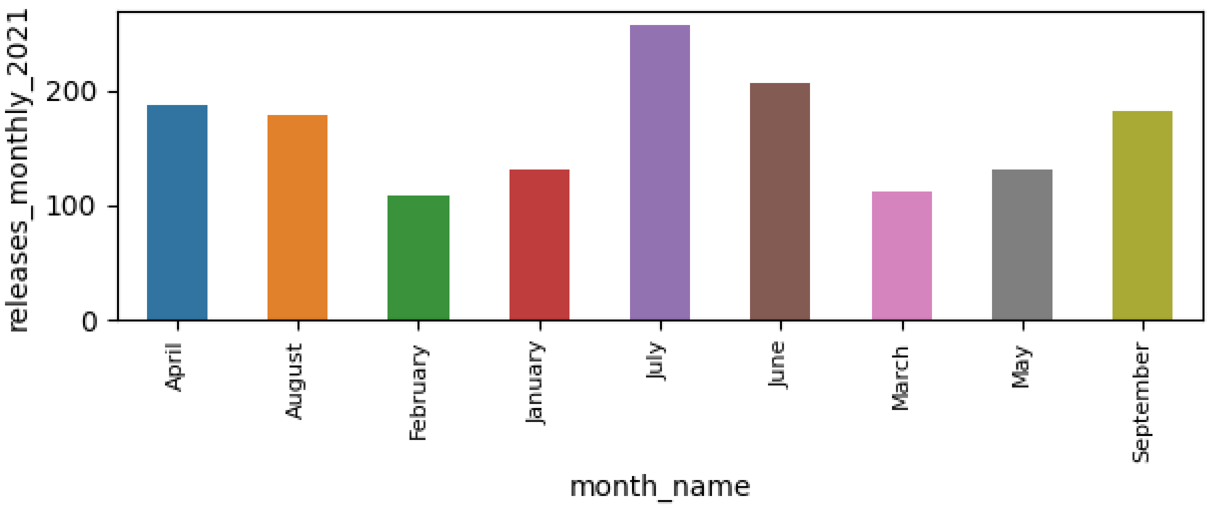
```
#4_a. Year 2020

data_2020 = data[data["year"] == 2020]
data_2020_monthly = data_2020.groupby("month_name")[["title"]].count().reset_index()
data_2020_monthly.rename({"title":"releases_monthly_2020"},axis = 1, inplace = True)
data_2020_monthly

plt.figure(figsize = (7,2))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_2020_monthly, x = "month_name", y = "releases_monthly_2020", width = 0.5)
plt.show()

#4_b Year 2020 and top countries

data2020 = data[["type","title","country","year","month_name"]]
data_2020_country = data_country.merge(data2020, on = "title", how = "left")
data_2020_country.drop("country_y",axis = 1,inplace = True)
data_2020_country.rename({"country_x":"country"},axis =1, inplace = True)
data_2020_country_1 = data_2020_country[(data_2020_country["year"] == 2020)
                                        & (data_2020_country["country"].isin(top_countries["country"].head()))]

plt.figure(figsize = (9,4))
sns.countplot(data = data_2020_country_1, x = "country", hue = "month_name")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()

#In 2020 also,content added throughout year.
#With respect to countries, Most of the US content has been added throughout year,followed by India.
#In December, a sharp increase of addition of content can be observed in United States.
```

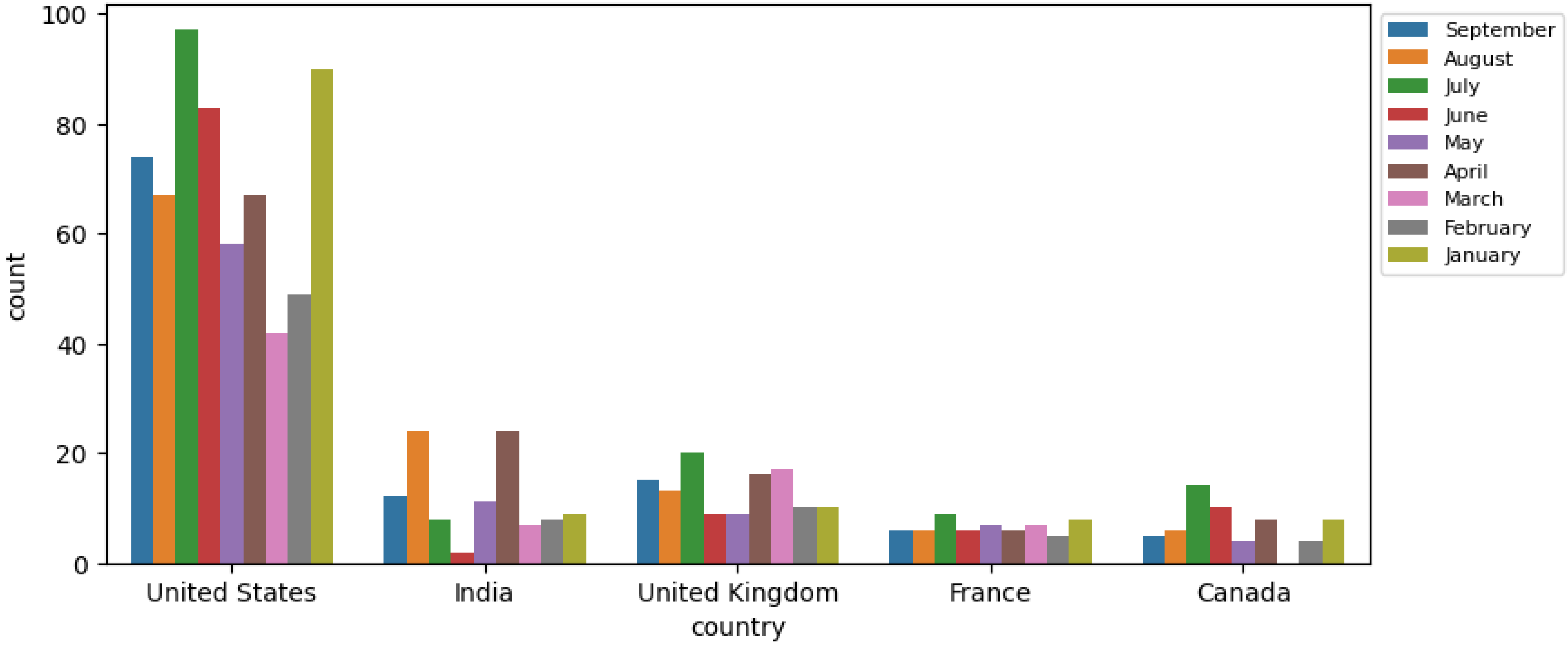

releases_monthly_2020
0
50
100
150
200
April
August
December
February
January
July
June
March
May
November
October
September
month_name


count
0
20
40
60
80
100
120
140
United States
India
United Kingdom
Canada
France
country
December
November
October
September
August
July
June
May
April
March
February
January

In [1384]:

```
#5_a. Year 2021

data_2021 = data[data["year"] == 2021]
data_2021_monthly = data_2021.groupby("month_name")[["title"]].count().reset_index()
data_2021_monthly.rename({"title":"releases_monthly_2021"},axis = 1, inplace = True)
data_2021_monthly

plt.figure(figsize = (7,2))
plt.xticks(rotation = 90, fontsize = 8)
sns.barplot(data = data_2021_monthly, x = "month_name", y = "releases_monthly_2021", width = 0.5)
plt.show()

#4_b Year 2021 and top countries

data2021 = data[["type","title","country","year","month_name"]]
data_2021_country = data_country.merge(data2021, on = "title", how = "left")
data_2021_country.drop("country_y",axis = 1,inplace = True)
data_2021_country.rename({"country_x":"country"},axis =1, inplace = True)
data_2021_country_1 = data_2021_country[(data_2021_country["year"] == 2021)
                                        & (data_2021_country["country"].isin(top_countries["country"].head()))]

plt.figure(figsize = (9,4))
sns.countplot(data = data_2021_country_1, x = "country", hue = "month_name")
plt.legend(bbox_to_anchor=(1, 1), loc=2, fontsize = '8')
plt.show()

#In 2021,a bit drop in adding content can be seen in Feburary and March but gradually back on track in July month
#The data for year 2022 is upto month September only.
#With respect to countries, Most of the US content has been added throughout year. Janaury and July are the
#peak months in US.
```

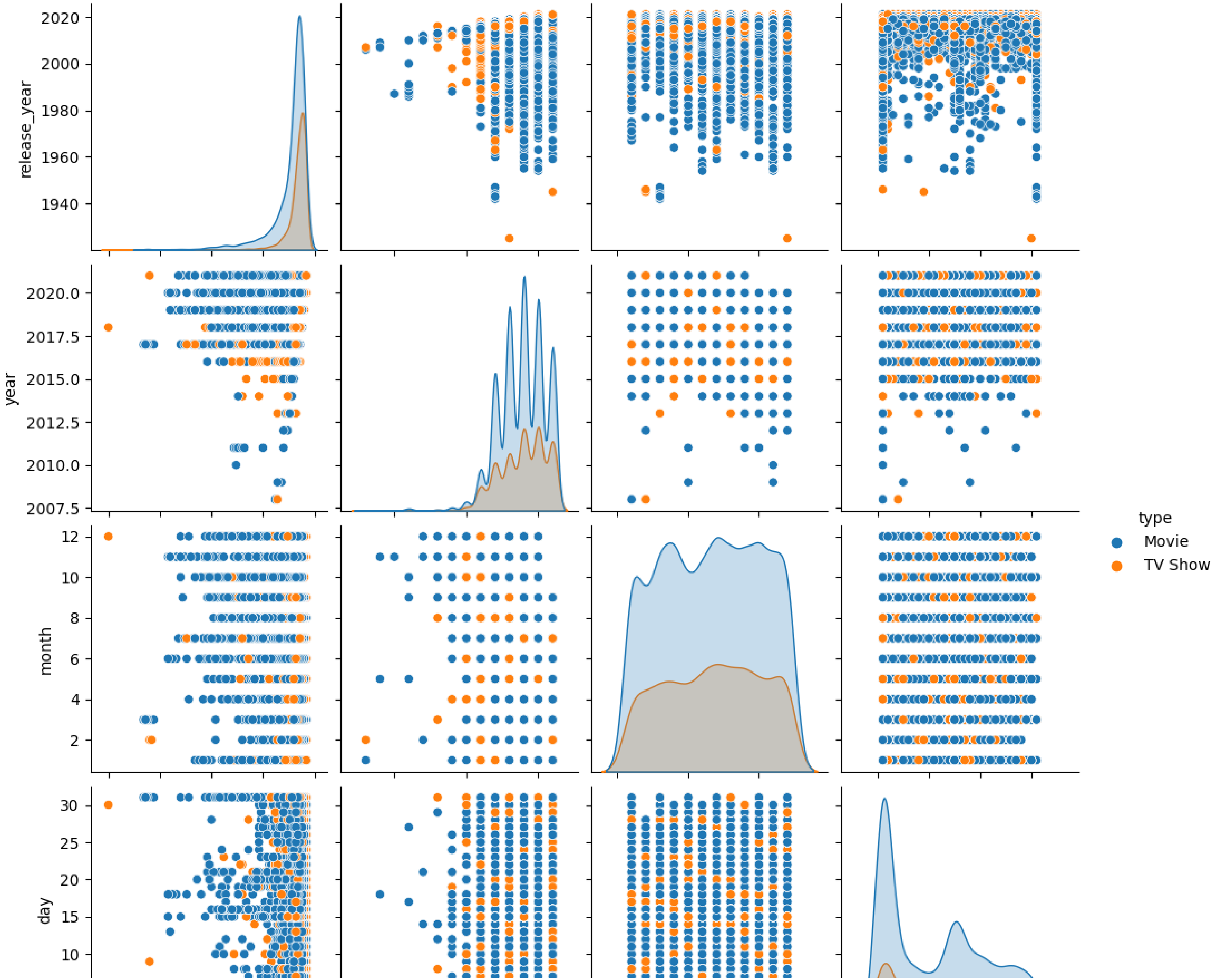

releases_monthly_2021
200
100
0
April
August
February
January
July
June
March
May
September
month_name


100
80
60
40
20
0
count
United States
India
United Kingdom
France
Canada
country
September
August
July
June
May
April
March
February
January

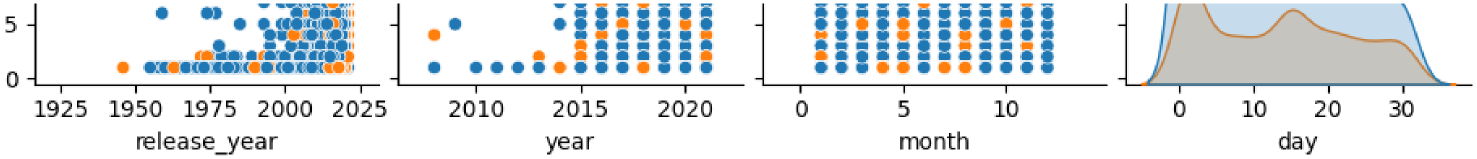
Other Analysis: correlation Analysis

In [1385]:

```
#correlation among different numeric data in dataset for movies and TV shows.

sns.pairplot(data = data, hue="type")
plt.show()

#pairplot will convey all the relation between all the columns.
```



release_year
2020
2000
1980
1960
1940
year
2020.0
2017.5
2015.0
2012.5
2010.0
2007.5
month
12
10
8
6
4
2
day
30
25
20
15
10
type
Movie
TV Show

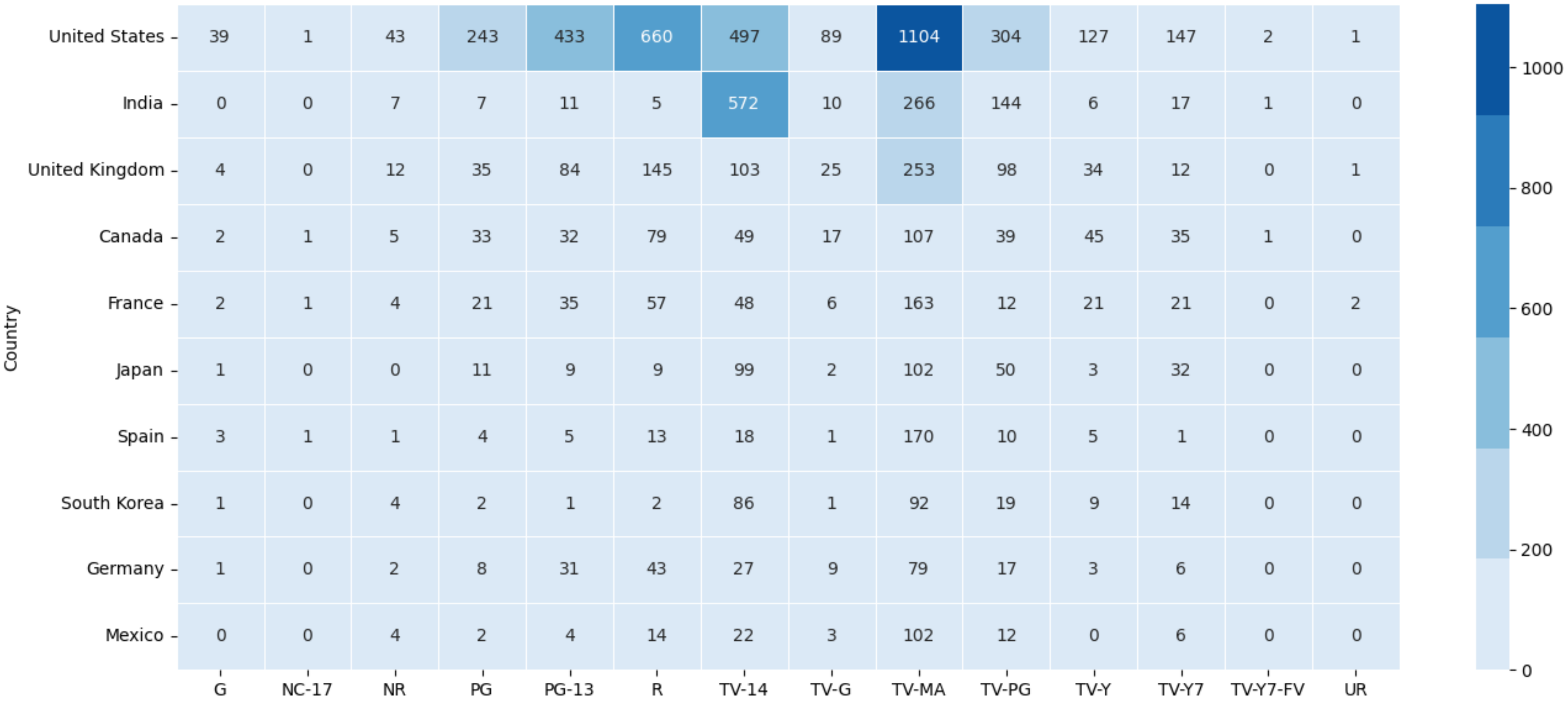


5
0
1925
1950
1975
2000
2025
release_year
2010
2015
2020
year
0
5
10
month
0
10
20
30
day

In [1387]:

```
# Top 10 Countries and ratings of Movies/TV shows

country_list = data['country'].apply(lambda x : str(x).split(',')).tolist()
country_list1 = pd.DataFrame(country_list, index = data['title'])
country_list1 = country_list1.stack()
country_list1 = pd.DataFrame(country_list1.reset_index())
country_list1.drop(['level_1'], axis = 1, inplace = True)
country_list1.rename(columns = {0:'Country'}, inplace = True)
country_list2 = country_list1.merge(data[['show_id', 'type', 'title', 'director', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description']], on = ['title'], how = 'left')
country_list2['Country'] = country_list2['Country'].str.strip()
country_count = country_list2['Country'].value_counts()
country_top10 = country_list2[country_list2['Country'].isin(list(country_count.index)[:11])]
country_top10 = country_top10[['Country','rating']]
country_top10['Count'] = 1

df1=country_top10.pivot_table(index="Country", columns="rating", aggfunc=["count"], fill_value=0)
column_names = []
for x in df1.columns:
  column_names.append(x[2])
df1.columns = column_names
df1 = df1.loc[['United States','India','United Kingdom','Canada','France','Japan','Spain','South Korea'
               ,'Germany','Mexico']]

colormap = sns.color_palette("Blues")
plt.figure(figsize=(16,7))
plt.title('Ratings for Movies/TV Shows over top countries')
sns.heatmap(df1 , annot = True, fmt="d", linewidths=.5,  linecolor='White',cmap = colormap)
plt.show()

#Heat map has been plotted for ratings of Movies/TV Shows across top Countries.
#United States has maximum ratings of TV-MA.
```

Ratings for Movies/TV Shows over top countries

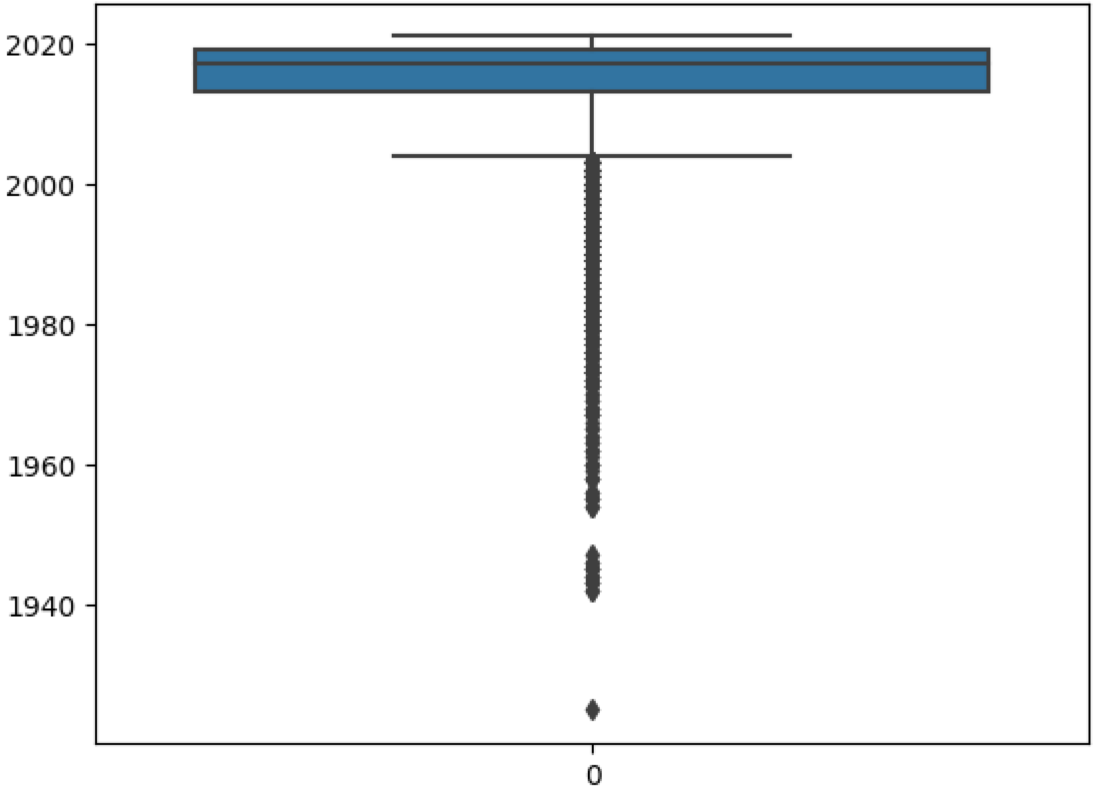


Country	G	NC-17	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y	TV-Y7	TV-Y7-FV	UR
United States	39	1	43	243	433	660	497	89	1104	304	127	147	2	1
India	0	0	7	7	11	5	572	10	266	144	6	17	1	0
United Kingdom	4	0	12	35	84	145	103	25	253	98	34	12	0	1
Canada	2	1	5	33	32	79	49	17	107	39	45	35	1	0
France	2	1	4	21	35	57	48	6	163	12	21	21	0	2
Japan	1	0	0	11	9	9	99	2	102	50	3	32	0	0
Spain	3	1	1	4	5	13	18	1	170	10	5	1	0	0
South Korea	1	0	4	2	1	2	86	1	92	19	9	14	0	0
Germany	1	0	2	8	31	43	27	9	79	17	3	6	0	0
Mexico	0	0	4	2	4	14	22	3	102	12	0	6	0	0

Q5

Missing Value & Outlier check (Treatment optional)

In [1388]:
```
# 1. Outliers in Release_year of Movies and TV shows present on Netflix Platform

sns.boxplot(data["release_year"])
plt.show()

# There are many outliers. Maximum number of movies added on Netflix platform have been released after 2000.
# The movies released before 2000 aren't contributing much in analysis or decision making.
```

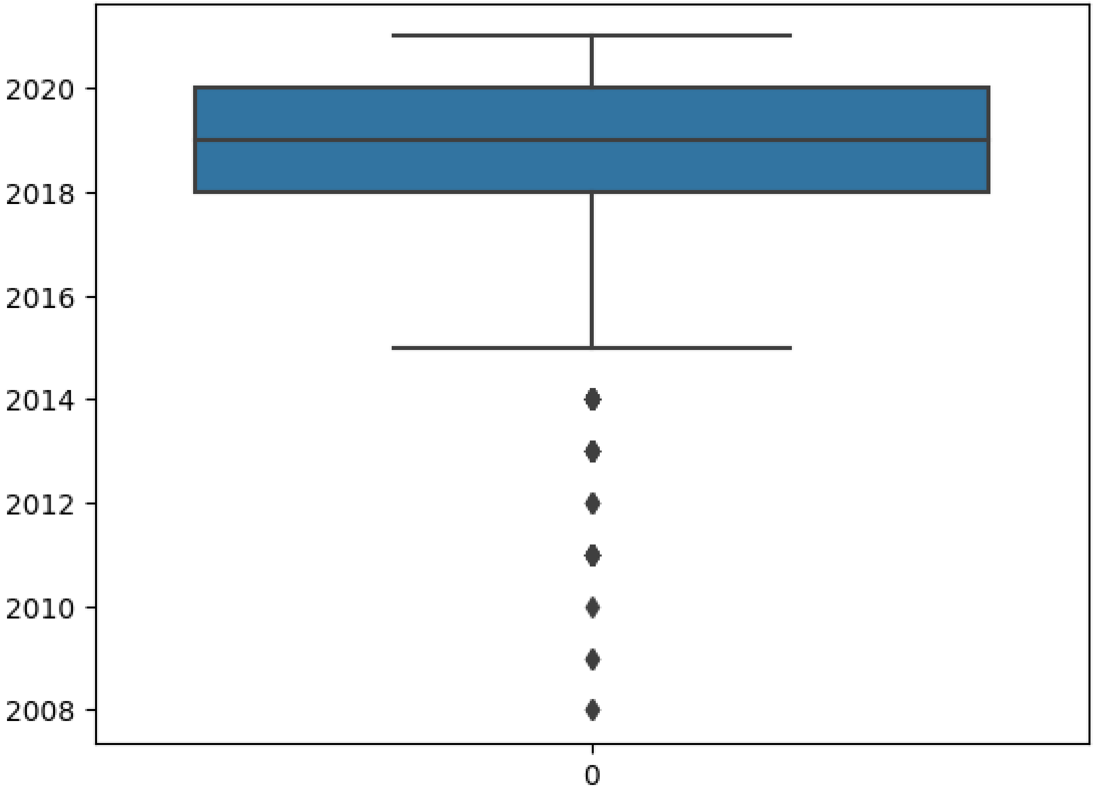


In [1390]:

```
# 2. Outliers in Years of Movies and TV shows added on Netflix Platform

sns.boxplot(data["year"])

# There are few outliers. Maximum number of movies added on Netflix after 2018, compared to previous years.
# It can be seen that movies added before 2015 aren't playing much role in future decision making as data
# is very less in these years and cannot be predict anything based on that data.
# However, if we observe, Movies and TV shows which are added after 2015 are useful in analysis of dataset
# and can be a deciding factors.
```

Out[1390]:

```
<Axes: >
```

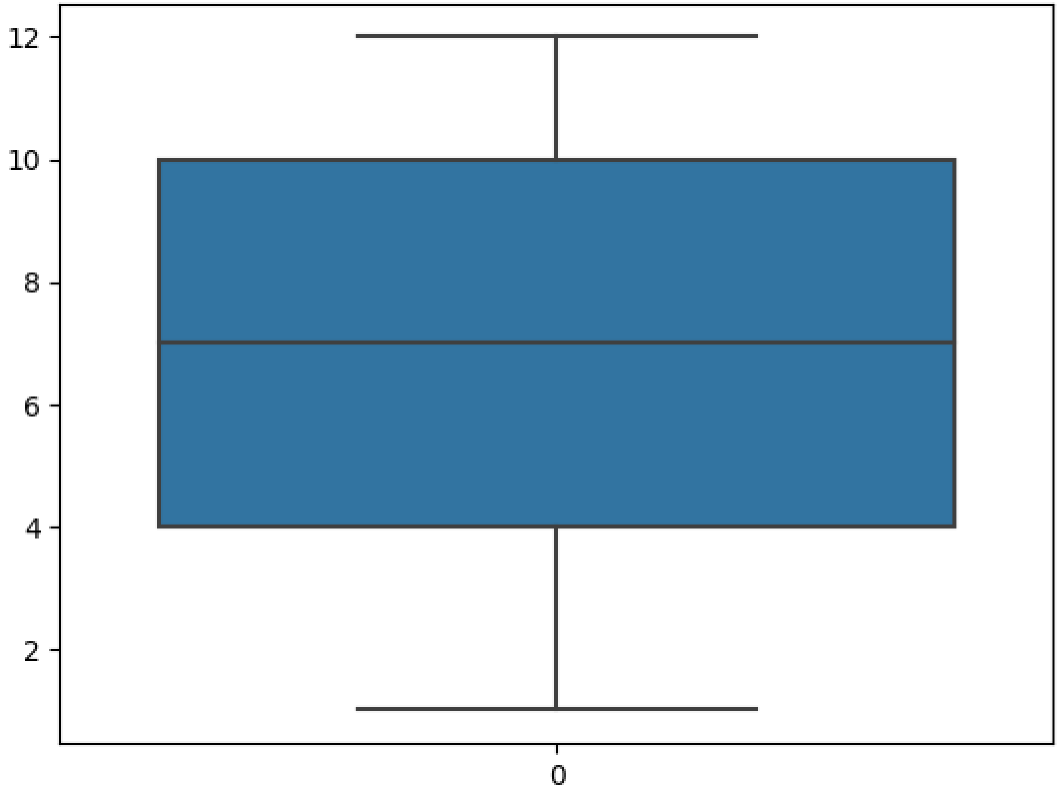


In [1391]:

```
# 3. Outliers in months of Movies and TV shows added on Netflix Platform

sns.boxplot(data["month"])

# There are no outliers in Months. All months are responsible for adding content on platform
# It can be seen that there is uniform distribution of months for movies added. However,a bit inclined
# towards few months in some specific countries.
```

Out[1391]:

```
<Axes: >
```



Q6

Insights based on Non-Graphical and Visual Analysis

6.1 Comments on the range of attributes

```
    All the comments have been attached with each code analysis. Please scroll above
    and refer all the codes.For quick summary, some points have been listed below

1. The dataset contains 8807 entries of movies/TV shows along with their 12 attributes.
2. There are total 12 attributes in dataset.
3. The attribute "release_year" is an integer type attribute.
4. Remaining attributes are object type.
5. There are multiple attributes which have missing values like director,cast, date and
   release year of movies/shows, rating and duration.
6. The maximum values are missing in director column.
7. There are 14 types of unique ratings in the dataset.

        Notice:It has been observed that the rating was categorized as ['PG-13', 'TV-MA', 'PG',
        'TV-14', 'TV-PG','TV-Y', 'TV-Y7','R','TV-G', 'G', 'NC-17', '74 min', '84 min', '66 min',
        'NR', nan,'TV-Y7-FV', 'UR'] and three entries for ratings('74 min','84 min','66 min')
        were not actually a category of ratings.
        There were null values for duration of three movies named "Louis C.K. 2017","Louis C.K.:
        Hilarious", and "Louis C.K.: Live at the Comedy Store". It has been concluded that the
        data of duration has been listed under column rating for these three movies. So, there
        was need of cleaning the dataset and same was cleaned for further processing.
```

6.2 Comments on the distribution of the variables and relationship between them

```
    All the comments have been attached with each code analysis. Please scroll above
    and refer all the codes.For quick summary, some points have been listed below

1. The first release happened in 1925 and latest release happened in 2021 according to given dataset.
   Alternatively, this is the time span for which dataset has been provided.
2. Netflix has first added their content on Janaury 1st, 2008 and the latest content is added on
```

```
   September 25th, 2021 according to dataset
3. The most releases of content have been happened in United States, followed by India and United Kingdom.
4. The most popular cast is Anupam Kher with 43 movies/TV shows enacted.
5. The most popular director is Rajiv Chilaka with 22 movies directed.
   Rajiv Chilaka has directed movies of series of "Chota-Bheem", which is very popular among kids.
6. The count of movies, that is 6131, is far greater than count of TV Shows,that is, 2676. People are
   fond of movies rather than TV shows.
7. The cinema has evolved a lot. Earlier, movie making is such a huge task. With advancements of technology
   and enhanced talent, most of people try their hands on production now. It can be observe from the plots
   that after 2000, the production of movies and TV shows have been increased drastically.
8. Netflix has started adding enormous amount of data from the last five years. From 2016, they have raised
   the quantity in their platform. The maximum movies/TV Shows has been added in year 2019.
9. It can be observed that maximum ratings have been provided under category "TV-MA", followed by "TV-14".
10. People love to watch International Movies followed by the Dramas.
```

6.3 Comments for each univariate and bivariate plot

```
    All the comments have been attached with each univariate and bivariate plot. Please scroll above and
    refer all the plots.
```

Q7

Business Insights - Should include patterns observed in the data along with what you can infer from it

```
    All the Insights have have been attached with each plot during analysis.
    For quick summary, some points have been listed below:

1. Netflix has attained popularity gradually. Netflix is in the market for last 12 years but for the last
   five years,it is attracting customers. Netflix has started adding huge content on their platform.
   In year 2016, the data added was far greater than the previous year and from then onwards, the count of
   added content is increasing.
   However, from 2019 to 2021, the Movie/TV Show content been added on their platform has shown a decline trend.
   And,the decline slope of movies is higher than slope of TV Shows.

2. The feburary and may month has less added data, whereas all other months have considerable same amount of
```

added content. The month July and December have maximum added content.
It can be observed that the Fall(month from July to December) season has large amount of content added compare to the remaining months.

3. With time, the interest in watching movies and TVshows has increasing only. However, since 2015, with the easily availability of internet and increasing usage of mobiles, the watching time has been increased abruptly.

4. Earlier, people fond of movies but audience of present time would love to watch not only movies but also TVshows.
   The creation of good TV shows like "Breaking Bad", "Game of Thrones", "Stranger Things" have diverted the prime time from movies to TV shows. TV shows are so much in demand that in the year 2021, the watching time of TV shows is greater than of movies.

5. Movies created are mostly of genres like International Movies,Dramas, and Comedies while TV shows are of genres like International TV Shows, TV Dramas, and TV Comedies.It can be inferred that people tend to like Content which is Internationally acclaimed, followed by Dramas and comedies.

6. People of most of the countries are likely to watch Movies compared to TV Shows. The countries like United States and India can be seen having huge difference between viewers of Movies and TV Shows.
   Japan and South Korea are two exceptions countries in top 20 countries that have more number of viewers in TV Shows than Movies.

7. The US content has got all types of ratings, followed by Canada. It is more unlikely to get ratings of "UR" and "TV-Y7-FV".

8. Popular directors are more likely to direct movies. There are some exceptions who try their hands on both type of content. "Marcus Raboy", "Anurag Kashyap", and "Ryan Polito" are such directors who explore area of TV shows also. However, comparison to movies count,number of TV shows directed by them are too less.

9. Popular actors are more likely to act in movies instead of TV Shows. The most popular actor "Anupam Kher" has acted in both movies and shows. Similarly, from top10 popular actors, only three actors(Anupam Kher, Julie Tejwani and Rupa Bhimani) have acted in both types of content.
   There are certain actors who have acted in TV shows and gain equal popularity. Takahiro Sakurai, Yuki Kaji, Ai Kayano, and others are such actors. It's important to note here that these actors are from Japan where TV shows are more popular than movies.
   Actors who are among famous top25 to top50 list, are equally devoted to both type of content.

10. Jay Karas and Marcus Raboy directed movies from United States have been attained the maximum place in Netflix platform.
    Similarly, Youssef Chahine from Egypt and Jan Suter from Mexico have directed more than 10 movies/shows and have been able to grab place in Netflix.
    Most Indian movies/TV Shows on Netflix platform have been directed by Anurag Kashyap and David Dhawan.

11. Anupam Kher from India has acted in 40 movies/shows followed by Shahrukh Khan. However,Shahrukh khan has only acted in movies.
    Takahiro Sakurai from Japan has acted in about 29 TV shows.
    Samuel L. Jackson from United States has acted in 22 movies/TV shows.

12. United States has a very versatile movie/shows production, followed by India and United Kingdom.
    United States have movies with maximum kind of genres.However, the most popular genre among US people is Dramas and Comedies.Some people prefer documentaries also in US.
    Indian people like movies of genre International Movies,Dramas, and Comedies. However, some selective Indian audiences prefer action and romantic movies also.
    People of France prefers only International Movies and Dramas.
    Japanese are much into TV Shows. The most popular genres in Japan are International TV Shows and Anime Series.
    Similar to Japan, South Koreans also prefer TV Shows. The most popular genres in South Korea are International TV Shows and Korean TV shows. The people are so interested in TV shows that the whole category of genre has been assigned to it.

13. "The Cathy Garcia-Molina and Joross Gamboa" of Phillipines, "David Dhawan and Anupam Kher" from India, and "Fernando AyllÃ³n and Ricardo Quevedo" from Spain & Colombia are some best Director-Actor pairs.

    Exceptions in terms of Actor-Director pair:
    S.S. Rajamouli had directed "Bahubali-The Beginning" and "Bahubali-The conclusion", but as this movie has been released in multiple languages, the count has been increased.
    Rajiv Chilaka had many casts on "Chota-bheem" movie series.
    Toshiya Shinohara from Japan has directed the series of Movie "InuYasha" which has many casts.

Q8

Recommendations-Actionable items for business.

Recommendations based on above Insights are:

1. Movies are an integral part of entertainment. However, based on last five year trend, TV shows are also attracting world-wide subscribers.
   Netflix should continue with the production and addition of movies on their platform.

Netflix should take risk for good concept TV Shows. It can brought maximum revenue. As TV Shows have duration in seasons,they bound their subscriber for long time. Good and new concept TV Shows helps in retaining subscribers.

2. US subscribers are likely to watch movies of all kind of genres. They watch movies more than TV Shows. Netflix can add good US movies to its platform. Similar action can be taken for Indian movies also.

   Japanese and South Koreans prefer TV Shows. Netflix should add TV shows for these countries. If it's in production, Netflix must prefer to produce TV Show instead of movies for Japanese and South korean audience.

3. All dircetors of US are good. However, Netflix can prefer Jay Karas and Marcus Raboy as they have directed maximum and good movies.
   Similarly, Netflix can approach Youssef Chahine from Egypt and Jan Suter from Mexicos.
   In Indian context, Anurag Kashyap and David Dhawan are good directors.

4. For Indian entertainment, Anupam Kher is a versatile actor and suited for both TV Shows and movies. If it is only movie, Netflix should connect with Shahrukh Khan.
   In US, Netflix can approach Samuel L. Jackson
   Takahiro Sakurai, Yuki Kaji, and Ai Kayano are famous among TV SHows. Netflix can cast Takahiro Sakurai from Japan in its TV Show production.

5. US people are likely to watch all kind of genres but the most poular among them is Dramas and Comedies.
   Netflix can focus on content with those genres.
   Japanese people love to watch TV Shows and Anime Series. Hence, if Netflix wants to try their hands in production of animation, they can prefer to produce and release content in Japan and South Korea.

6. Netflix can prefer Director-Actor pair "The Cathy Garcia-Molina and Joross Gamboa" of Phillipines,
   "David Dhawan and Anupam Kher" from India, and "Fernando AyllÃ³n and Ricardo Quevedo" from Spain & Colombia.

Answers to asked Questions

1. How has the number of movies released per year changed over the last 20-30 years?

The number of movies released has been drastically increased over last 20-30 years.

2. Comparison of tv shows vs. movies.

Movies are more popular than TV Shows. Except Japananese and South Koreans, mostly all audience

```
    prefer watching movies.
    However, over last few years people are also inclining towards TV Shows.

**3. What is the best time to launch a TV show?**

    Feburary and May month has less added data, whereas all other months have considerable same amount
    of added content. The month July and December have maximum added content.
    It can be observed that the Fall(month from July to December) season has large amount of content
    added compare to the remaining months. Therefore, best time to launch a TV Show is Fall

**4. Analysis of actors/directors of different types of shows/movies.**

    Refer Points (8-11) and 13 of Q7- Business Insights

**5. Does Netflix has more focus on TV Shows than movies in recent years**

    Netflix has always focusing on movies but Yes, in recent years, along with movies, they also started
    focusing on TV Shows also.

**6. What content is available in different countries**

    US and India have more movies than TV Shows. But trend is little changing. There is also increase in
    count of TV Shows since last few years. They prefer Dramas and Comedies
    Japan and South Korean have more TV Shows than movies. They prefer Shows and Animes
    UK,Canada and France are among top countries and having movies more than TV Shows.
```